第 42 卷 第 6A 期
2015 年 6 月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

基于深度神经网络的汉语语音合成的研究

王 坚 张媛媛

(中央财经大学信息学院 北京 100081)

摘 要 为了提高基于 HMM 的语音合成的音质,探讨了不同的结构和参数对深度神经网络(DNN)训练的影响,并证明了 DNN 判别 S/U/V 的有效性,完成了 DNN 对 HMM 合成系统的合成语音谱参向原始语音进行转换。进一步地,探讨了对暂时分解(TD)算法得到的参数进行转换的方案,对 TD 分解得到的事件向量进行 DNN 训练,建立转换模型,并同未转换的事件函数进行再合成。实验证明,用 DNN 转换合成后的频谱更接近原始频谱;主观评测表明,该方法能有效地改善合成语音的音质。

关键词 HTS,DNN,深度学习,声音转换,暂时分解

中图法分类号 TN912.33 文献标识码 A

Title Research on Deep Neural Network Based Chinese Speech Synthesis

WANG Jian ZHANG Yuan-yuan

(School of Information, Central University of Finance and Economics, Beijing 100081, China)

Abstract In order to improve the quality of speech synthesis based on HMM, this paper discussed the different structure and parameters on the effect of DNN training and demonstrated the validity of DNN discriminating S/U/V. The paper finished the speech synthesis of DNN on the HMM synthesis system were converted to the original speech spectrum parameter. Then, we studied on temporal decomposition(TD) algorithm to get the parameters of conversion program, and for DNN training set up the conversion model and event with no conversion function resynthesis of event vectors. The experiment proves that DNN conversion spectrum synthesis is closer to the original spectrum, and the subjective evaluation shows that this method can effectively improve the synthesized speech quality.

Keywords HTS, DNN, Deep leaning, Voice conversion, Temporal decomposition

1 引言

语音处理技术是以语音语言学和数字信号处理作为基础的一门综合性学科[1]。语音合成技术的日益成熟,一方面使其在人们的生活中得到了越来越广泛的应用,而另一方面也使人们对语音合成系统的要求越来越高[2]。基于隐马尔可夫模型(HMM)的统计参数语音合成技术因为其较优秀的合成效果,且便于通过对模型参数的调整达到声音转换的目的,成为了目前最受关注的方法之一。然而 HMM 合成声音仍然存在声音过于平滑、沉闷、缺乏细节、自然度不高等影响音质的问题需要解决[3]。本文为了提高基于 HMM 的语音合成的音质,用少量的数据,从参数转换的角度,运用深度神经网络(DNN)对不同的参数进行训练得到转换模型,重新合成达到提升合成音质的效果。

2 隐马尔可夫模型的语音合成系统

将隐马尔可夫模型对语音参数进行建模用于语音合成的系统,是基于统计参数建模的语音合成研究和使用最广泛的方法之一。隐马尔可夫模型是一个双内嵌式随机过程,一个随机过程描述状态的转移,这个与语音中的声学参数的变化具有相似性,短时平稳且隐含不可观测,需要通过可观测序列来估计声学参数进行合成[4];另一个随机过程描述状态和观察值之间的对应关系,这恰好可以模拟可观测的语音信号序列和隐藏在这之下的合成参数的一个对应关系。因此,HMM 很符合人类的语音产生机理,是一种很适合进行语音信号分析处理的模型,其整个合成系统框图如图 1 所示。

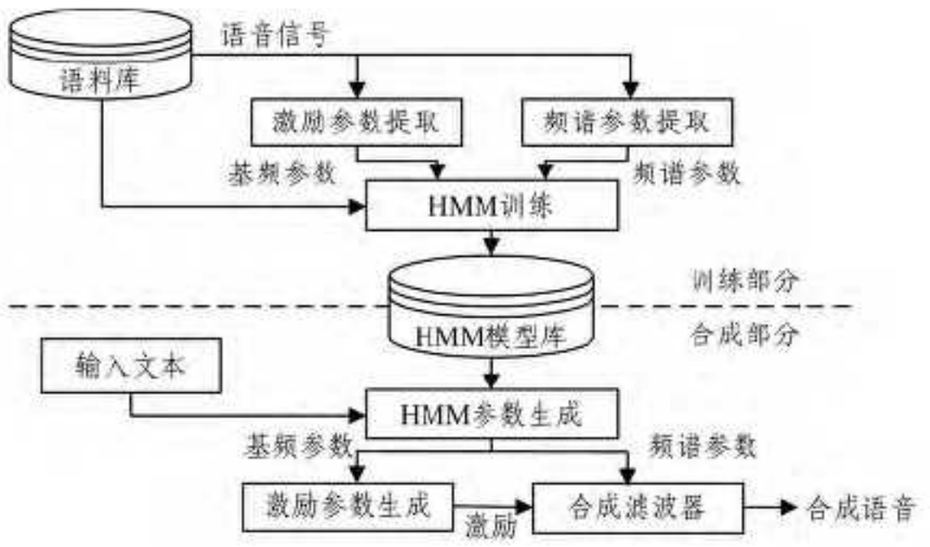

图 1 基于隐马尔科夫模型的语音合成系统

HTS 的训练部分的作用就是由最初的原始语料库经过处理和模型训练后得到这些训练语料的 HMM 模型[5]。建

本文受中央财经大学重点学科建设项目,北京高等学校青年英才计划项目(YETP0988)资助。

王 坚(1975—),男,博士,讲师,主要研究方向为模式识别、数据挖掘,E-mail:wanderingful@126.com;张媛媛(1982—),女,硕士,工程师,主要研究方向为数据挖掘。

模方式的选择首先是状态数的选择，因为语音的时序特性，一个模型的状态数量将影响每个状态持续的长短，一般根据基元确定。音素或半音节的基元，一般采用5状态的HMM；音节的基元一般采用10个状态。在实际的建模中，为了模型的简化，可以将HMM中的转移矩阵用一个时长模型(dur)替代，构成半隐马尔可夫模型。用多空间概率分布对清浊音段进行联合建模，可以取得很好的效果。

HTS的合成部分相当于训练部分的逆过程，作用在于由已经训练完成的HMM在输入文本的指导下生成参数，最终生成语音波形。具体的流程是：(1)通过一定的语法规则、语言学的规律得到合成所需的上下文信息，标注在合成label中。(2)待合成的label经过训练部分得到的决策树决策，得到语境最相近的叶结点HMM就是模型的决策。(3)由决策出来的模型解算出合成的基频、频谱参数。根据时长的模型得到各个状态的帧数，由基频、频谱模型的均值和方差算出在相应状态的持续时长帧数内的各维参数数值，结合动态特征，最终解算出合成参数。(4)由解算出的参数构建源_滤波器模型，合成语音。源的选取如上文所述：对于有基频段，用基频对应的单一频率脉冲序列作为激励；对于无基频段，用高斯白噪声作为激励。

3 基于深度神经网络的语音合成

本文根据深度神经网络的特点，针对语音合成的现状和HTS本身的缺点，提出一种用深度神经网络进行参数转换的方法来改进HTS的合成效果[8]。只要通过合适的转换，一个有效训练的模型就能够合成出各种音色的声音[6]。

3.1 参数转换语音合成策略

将合成语音和原始语料看作两个独立的说话人，其参数分别作为源、目标向量，即用深度神经网络表示一个从合成语音到原始语料的整体参数映射关系。

LSF具有按阶排列的特性：

$$0<\omega_1<\theta_1<\omega_2<\cdots<\omega_{p/2}<\theta_{p/2}<\pi \tag{1}$$

相邻参数的差总大于零，在参数密集的地方表示在该频率段存在一个共振峰，而在参数稀疏的地方表示一个低谷，这样也能直观地表示频谱分布，同时保证LPC合成滤波器的稳定性[7]。实验考虑使用LSF参数作为训练参数。依照HTS的建模基元，以及现有训练语料的规模，在此使用音节为单元进行映射网络模型，就是对源_目标的每一个音节(单字)建立一个深度神经网络进行转换。

将同维的参数进行归一化，扩大帧与帧之间参数的差异性。与之前不同的是，在深度神经网络进行参数转换后，转换的参数还需要经过反归一化还原用于合成转换语音。本文只选取与目标参数最接近的20组数据进行转换，并使用了一个转换/替换的方案，当转换的参数不符合按阶排序时，就用目标语料中距离最近的向量直接替换该参数，以达到强制转换的目的。另外增加一个自适应步长的学习效率调整加快网络的整体调优，防止因学习效率不合适而难以收敛到极小值。具体方案是在初始学习效率的基础上，每调优一次，比较两次之间的误差距离，如果调优后的误差距离超过之前的某个数值(这里根据经验设为1.03)，就将学习效率降低至原来的0.9倍。

3.2 参数转换语音合成架构

该系统的架构可以分为网络训练阶段和转换合成阶段两个部分，网络训练阶段框图如图2所示。

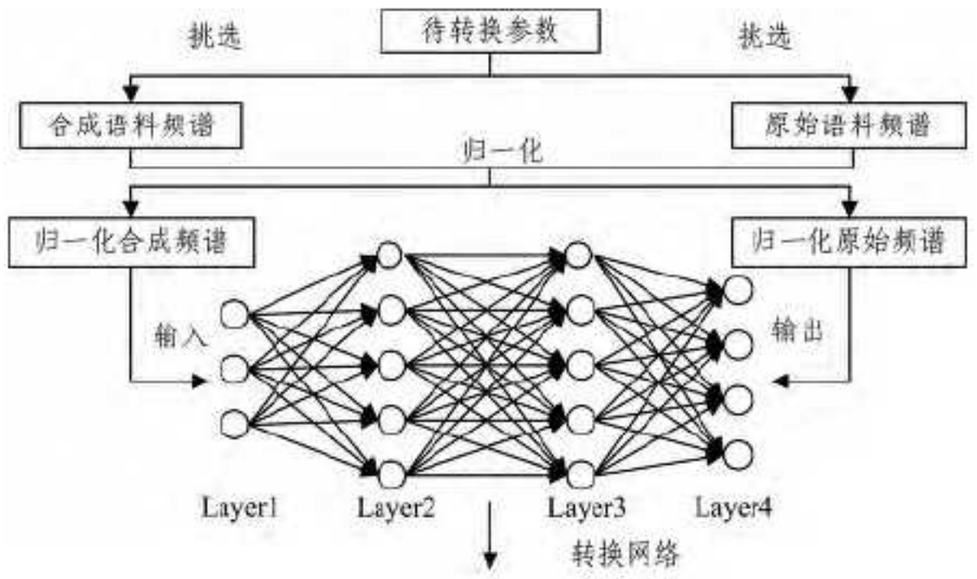

图2 参数转语音合成系统训练阶段框图

首先由待转换参数的标注信息，从合成语音和原始语料中选取相同模型单元(音节)的平行语料的频谱参数，经过时间对齐后统一按维进行归一化处理，得到的归一化参数分别作为深度神经网络的输入参数和输出参数进行学习。得到各个音节的源_目标转换网络。

其转换合成阶段的框图如图3所示。

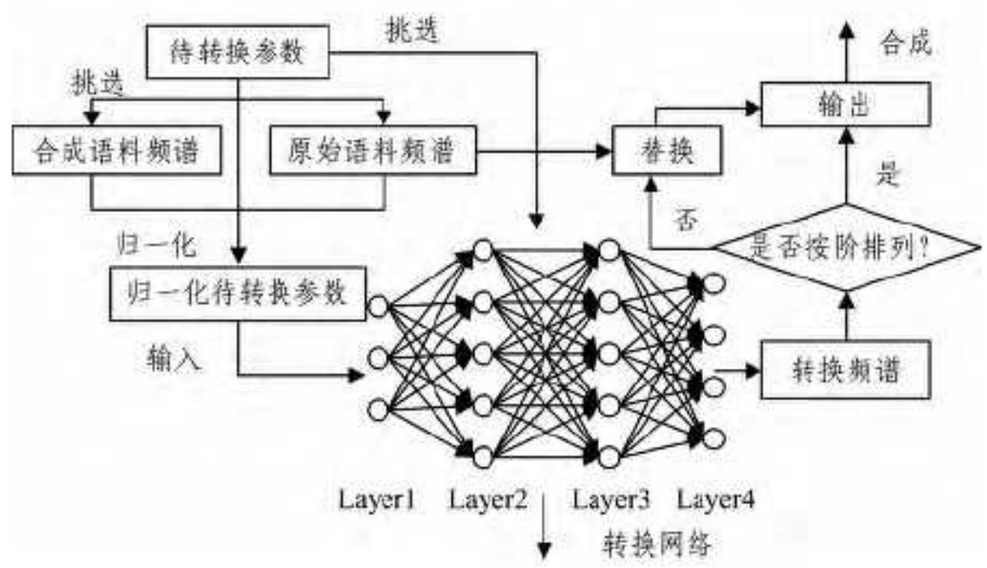

图3 参数转换语音合成系统转换合成部分框图

首先由待转换参数的标注信息挑选出对应的深度神经网络用于转换，同时挑选出合成语料和原始语料的频谱用于统一的归一化处理。归一化后的待转换频谱就作为对应转换网络的输入参数，经过深度神经网络输出得到转换后的频谱。判断转换频谱是否具有按阶排列的特性，如果不具有则用原始语料的频谱替代此帧，最终得到稳定的转换频谱，最后通过滤波器合成。这样，在有限的训练参数的条件下也能够完成训练、转换并合成的步骤，达到提高音质的效果。

3.3 暂时分解参数转换语音合成

基于深度神经网络的暂时分解参数转换语音合成系统框图如图4所示。

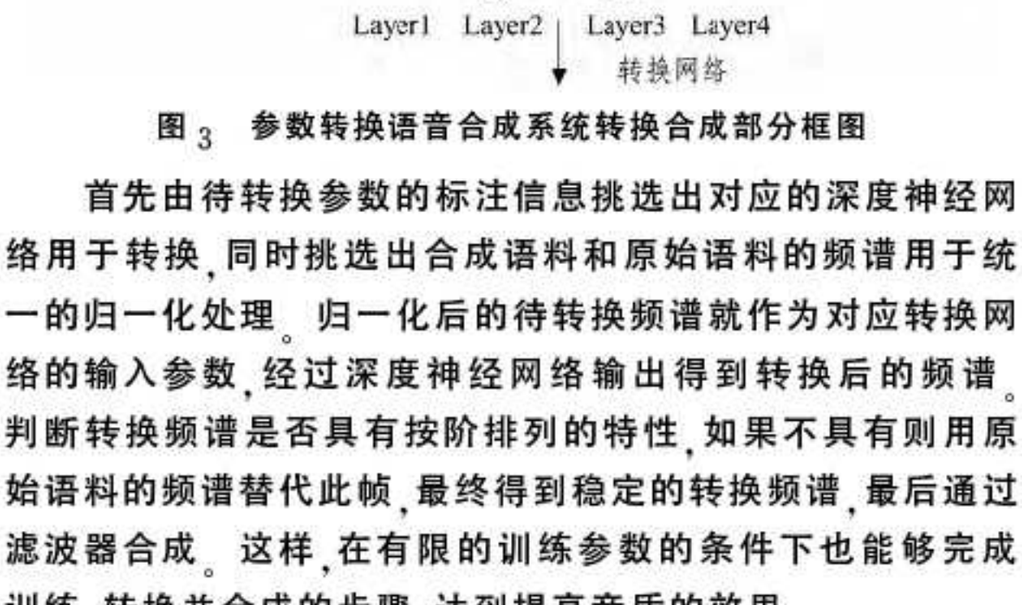

图4 暂时分解参数转换语音合成系统转换合成阶段框图

首先将待转换的频谱参数暂时分解得到事件向量和对应的事件函数。保持事件函数不变，将事件向量作为转换特征，根据所处转换模型挑选出之前训练用的原始语料和合成语料的平行语料频谱参数，进行统一的归一化处理；同时挑选出相应的转换深度神经网络。最后将转换得到的事件向量与最开始的事件函数重新合成出转换的频谱参数[9]，再与对应的基频(f_0)和非周期特征(ap)一起合成出转换的语音。

4 实验验证与分析

4.1 静音/清音/浊音(s/u/v)的判别实验

采用深度神经网络的方法进行$s/u/v$的判别实验。实验数据来自实验室录制的电话语音库，采集于市话信道，语音采样率为$8kHz,16bits$量化。选取了其中一个说话人的一共5000帧语料，帧长$10ms$，帧间无重叠。考虑到清浊音段的参数变化多于静音段，3种语料的数量为：静音1000帧、清音2000帧、浊音2000帧。

4.1.1 DNN的结构对训练结果的影响

采用按特征维归一化的思想，即对所有帧的特征维单独归一化，归一化函数为：

$$\hat{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

其中，$\hat{x}_i$和x_i分别是归一化后和归一化前该维的特征，$x_{\max}$和$x_{\min}$分别是该维特征的最大、最小值。深度神经网络参数为：$batchsize$(数据组大小)为100，$dnn\_numepochs$(隐层学习迭代次数)为5，$bp\_numepochs$(整体调优迭代次数)为20，α(学习效率)为1。实验采用的方法是在实验完i层隐藏层后，统计它们的识别率，取最佳结构(指识别率处于前10的结构)作为增加一层隐藏层的前i层结构，依次设置第$i+1$层隐藏层神经元数目$1-200$，比较其识别率，实验结果如图5所示。

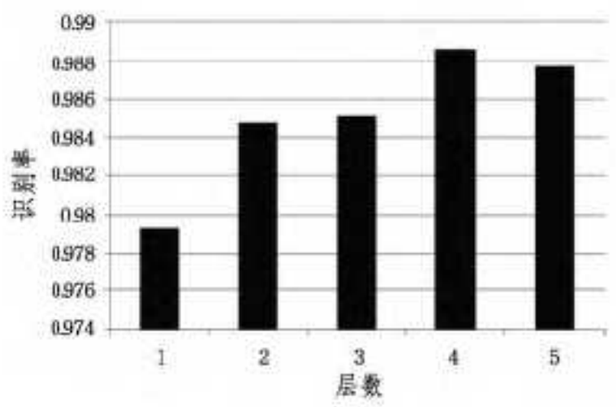

图5 各层神经网络识别率对比

由图5的各层最佳结构深度神经网络得到的平均识别率可以看出，一层隐藏层的神经网络即传统的神经网络的识别率一般，而随着隐藏层的增加，识别率都稳步上升，但五层深度神经网络在识别率上反而有所下降。

4.1.2 DNN的参数特征对训练结果的影响

深度神经网络参数设置如表1所列。

表1 深度神经网络参数设置

网络结构	100	80	80
batchsizes	100	α	1
dnn-numepochs	5	bp-numepochs	100

此处选取了一个相对随机的结构作为实验结构，测试识别率在特征数量变化下的改变情况，实验结果如图6所示。

由图$6(a)$可以看出，随着训练特征数量的增加，识别率提升到稳定点的速度也越来越快。图$6(b)$是$300-3000$帧训练特征前100次迭代的识别率的平均值，用以表示相应训练条件下的平均识别率。可以看出，随着训练特征数量的增加，除了在$2100-2700$帧时识别率有一点下降外，总体的识别率是不断提高的，下降的原因经分析可能是$2100-2700$帧的语料特征不太明显，导致深度神经网络在学习结果性能上的一些下降。推出结论：训练特征的数量增加能加快深度神经网络学习的收敛，并提高深度神经网络学习的效果。

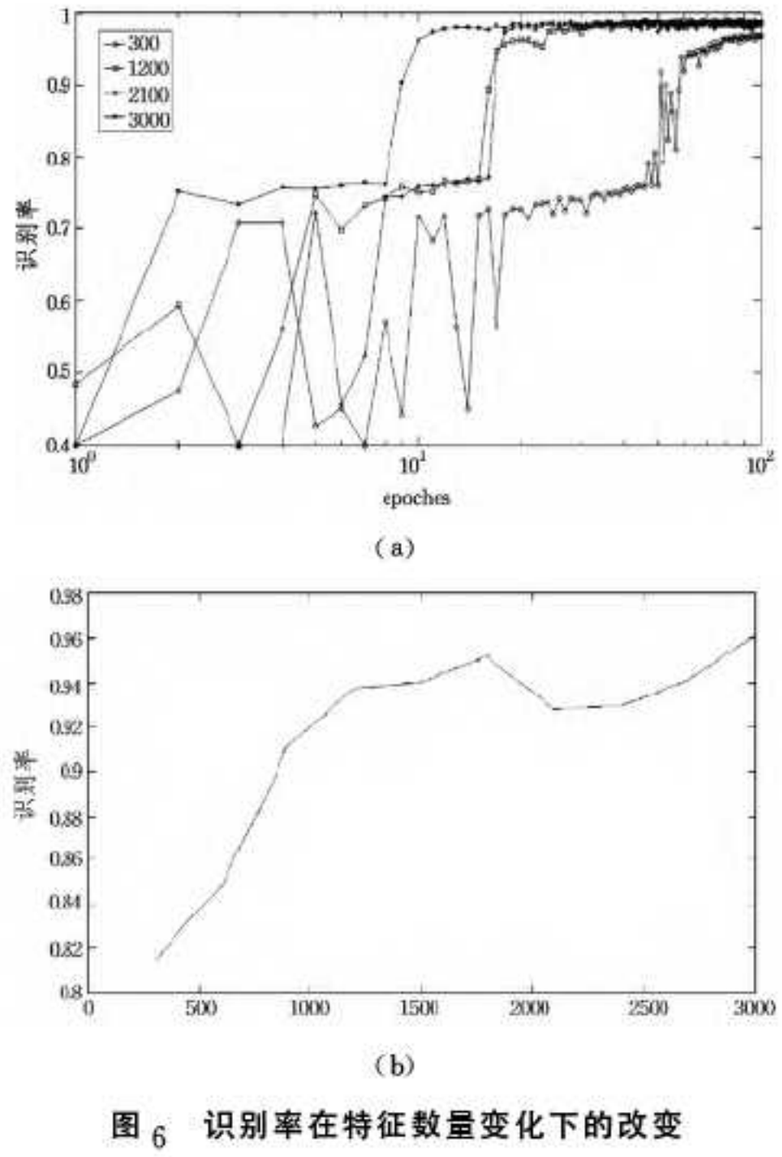

图6 识别率在特征数量变化下的改变

4.1.3 DNN进行$s/u/v$的判别

将全部5000帧语料按照$s/u/v$比例分为5组，每组1000帧。实验时轮流随机抽取其中的3组作为训练语料，其他2组作为测试语料，实验5次。参数设置如表2所列，识别率结果如图7所示，误识率统计如表3所列。

表2 深度神经网络参数设置

batchsizes	100	α	1
dnn-numepochs	5	bp-numepochs	1-100

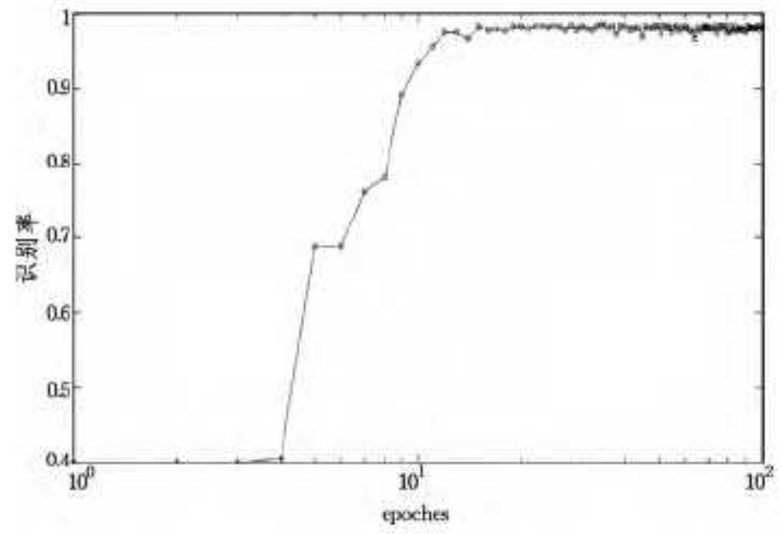

图7 深度神经网络对$s/u/v$的判别识别率随迭代次数的变化

表3 $s/u/v$误识率统计表

错误率(%)	→s	→u	→v
s	—	0.0098	0
u	0.0735	—	1.4985
v	0	2.9887	—

从图7可以看出，该深度神经网络在迭代大约13次后达

到稳定值，识别率为 98.3%。由表 3 可以看出，静音与清浊音之间互相识别错误率很低，而清音识别成浊音的错误率大约为 1.5%，浊音识别成清音的误识率大约为 3%，其原因可能在于某些清音在发音上的浊化或某些能量较小的浊音段等本身容易相互混淆，且清浊与浊音的交界处不能完全准确地判断可能导致了标注信息中的一部分错误。总的来说，实验结果证明了深度神经网络用于 s/u/v 判别的有效性。

4.2 DNN 参数转换语音合成实验

该实验在于验证基于深度神经网络的参数转换对于频谱参数如 LSF 的有效性，并合成验证其能提高合成语音音质。首先根据测试语料的标注在训练语料和合成语料中匹配对应的音节信息，并取出相应语料的对应音节所在帧区间的参数 LSF，得到各音节对应的平行语料组 dnn-in、dnn-out 用于训练深度神经网络，每组平行语料的帧数为 200—5000 不等，然后进行深度神经网络的训练。实验深度神经网络参数设定如表 4 所列。

表 4 深度神经网络参数设置

网络结构	80	60	60
batchsizes	10	α	1
dnn-numepochs	10	bp-numepochs	500

实验结果发现，16 维参数的转换效果在主观评测上的表现并不是十分理想，原因在于维度过少导致频谱细节体现不足，同时在再次合成的过程中的失真进一步加大；48 维的参数转换略好于 16 维，主要还是在于频谱细节转换得更多，但另一方面码本映射的影响导致的声音的不稳定也在一定程度上影响了听感。高 24 维的转换效果较好，在保留了低维部分后声音的自然度有较大改善。总体来说，该结果表明了基于深度神经网络的参数转换语音合成系统能有效提高合成的音质。

4.3 事件向量转换的 DNN 参数转换的语音合成实验

暂时分解的参数分别选用 16 维和 48 维的 LSF 参数，并同样进行针对 48 维 LSF 高 24 维的特征转换实验。先将测试语料的 LSF 参数进行暂时分解，得到一组事件向量 a_0 和一组事件函数 $\Phi_0(n)$，平均组数为 40 组/秒。保留事件函数不变，根据标注信息，挑选出每一组事件向量所对应的转换网络，并将转换参数与训练语料和合成语料中对应的参数统一归一化，作为深度神经网络的输入，输出转换参数经过转换/替换过程后再进行反归一化和得到最终的转换事件向量，与之前保留的事件函数重新合成 LSF 参数，再用 SPTK 工具包从 LSF 参数中重新提取出频谱参数 sp，与之前保存的基频 (f_0)、非周期特征 (ap) 一起合成转换后的语音。

客观评价采用相关系数，比较转换前后的 LSF 参数与训练语料 LSF 参数的相关系数。主观评价同一实验采用 A/B 偏好测试。其主观评测结果如图 8 所示。

由图 8 可以看出，使用事件向量进行的转换在主观评测上的得分都高于直接进行频谱参数转换的得分，这也证实了暂时分解中我们没有转换的事件函数能保留语音的可懂度，而转换的事件向量改善了频谱细节部分，提高了音质。在暂时分解的 3 个结果中，16 维的结果相对较差，原因还是在于维度相对少导致的频谱表现不够且失真较大，48 维的表现最好，而高 24 维的表现和 16 维的差不多，原因应该是暂时分解在一定程度上减少了直接码本映射导致的音质下降。总体来说，基于深度神经网络的暂时分解参数转换语音合成能更好地提升合成音质。

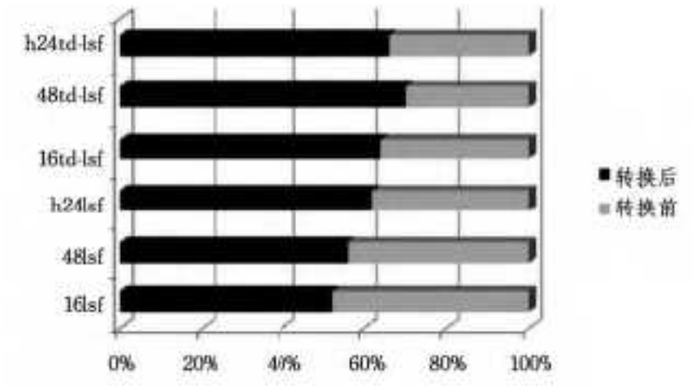

图 8 各参数转换主观评测结果

结束语 为了提高基于统计参数训练建模的 HTS 的合成效果，本文结合深度神经网络在语音合成中的应用，从参数转换的角度出发，提出一种新的合成系统，并通过实验证实了该系统的有效性，为 HTS 系统的发展以及多样化、深度神经网络及其在语音合成方面的研究奠定了基础。实验表明，该系统在性能上优于直接对频谱参数的转换，并能有效提高音质。由于本文采用的是一种码本映射的方法，该方法的缺点在于转换结果可能会出现不稳定等因素，因此在下一步的研究中，还需考虑进行其它如模型参数的转换，以弥补码本映射的不足。

参 考 文 献

[1] 井晓阳，罗飞，王亚棋. 汉语语音合成技术综述[J]. 计算机科学，2012，39(Z3)：386-391

[2] 赵鸿图，刘云. 改进粒子群算法的小波神经网络语音去噪[J]. 计算机测量与控制，2013，21(10)：2799-2802

[3] 赵建东，高光来，飞龙. 蒙古语语音合成语料库标注规则的设计[J]. 内蒙古大学学报：自然科学版，2013，44(3)：51-55

[4] 胡郁，凌震华，王仁华，等. 基于声学统计建模的语音合成技术研究[J]. 中文信息学报，2011，25(6)：275-279

[5] 宋阳. 基于统计声学建模的单元挑选语音合成方法研究[D]. 合肥：中国科学技术大学，2014

[6] 赵力. 语音信号处理(第 2 版)[M]. 北京：机械工业出版社，2011

[7] 孙志军，薛磊，许阳明，等. 深度学习研究综述[J]. 计算机应用研究，2012，29(8)：2806-2810

[8] Nandasena A，Nguyen P C，Akagi M. Spectral stability based event localizingtemporal decomposition[J]. Computer Speech and Language，2011，15(4)：381-401

[9] 殷力昂. 一种在深度结构中学习原型的分类方法[D]. 上海：上海交通大学，2012

(上接第 47 页)

[9] Wang C，Feng X J，Li X，et al. Colored Petri net model with automatic parallelization on real-time multicore architectures[J]. Journal of Systems Architecture，2014，60：293-304

[10] Santana-Robles F，Medina-Marin J，Montano-Arango O，et al. Modeling and simulation of textile supply chain through colored Petri nets[J]. Intelligent Information Management，2012，4(5A)：261-268

[11] 马良荔，陈杰，汪丽华. 模糊有色 Petri 网的形式化推理算法研究[J]. 计算机科学，2012，39：256-258