Science Discovery
2018; 6(4): 298-305
http://www.sciencepublishinggroup.com/j/sd
doi: 10.11648/j.sd.20180604.21
ISSN: 2331-0642 (Print); ISSN: 2331-0650 (Online)

Fundamental Frequency (F0) Fusion Transformation-Based on BLSTM for Voice Conversion

Miao Xiaokong, Zhang Xiongwei, Sun Meng

Command & Control Engineering College, Army Engineering University, Nanjing, China

Email address:
miao_xk@163.com (Miao Xiaokong), xwzhang9898@163.com (Zhang Xiongwei), sunmengccjs@163.com (Sun Meng)

To cite this article:
Miao Xiaokong, Zhang Xiongwei, Sun Meng. Fundamental Frequency (F0) Fusion Transformation-Based on BLSTM for Voice Conversion. *Science Discovery*. Vol. 6, No. 4, 2018, pp. 298-305. doi: 10.11648/j.sd.20180604.21

Received: June 25, 2018; **Accepted**: August 7, 2018; **Published**: August 10, 2018

Abstract: For the current speech conversion algorithms based on neural networks, the method of using the mean-variance linear transformation fundamental frequency (F0) can easily cause some "mechanical tones" and strange adjustments in the converted speech, and the similarity of the transformed speech is low. This paper proposes the nonlinear mapping of F0 using BLSTM (Bi-directional Long Short Term Memory) neural network, and merging the structural information with the source fundamental frequency. Firstly, the stable BLSTM network is trained through the paired fundamental frequency F0, and then the final required fundamental frequency is obtained by fusing the converted F0′ and the original structural information F0, and finally the speech synthesis is performed, thereby improving the similarity between the converted speech and the target speech. degree. At the same time, it is verified that the method proposed in this paper can reduce the mechanical sounds of speech conversion to a certain extent and improve the similarity of the converted speech.

Keywords: Voice Conversion, BLSTM Neural Network, Fundamental Frequency Conversion, Nonlinear Mapping

基于BLSTM实现基频(F0)融合变换的语音转换方法研究

苗晓孔，张雄伟，孙蒙

指挥控制工程学院，陆军工程大学，南京市，中国

邮箱
miao_xk@163.com（苗晓孔），xwzhang9898@163.com（张雄伟），sunmengccjs@163.com（孙蒙）

摘要：针对目前基于神经网络的语音转换算法中，采用均值方差线性变换基频(F0)的方法易造成转换语音中存在部分"机械音"和怪调以及转换的语音相似度低等问题。本文提出了采用BLSTM(Bi-directional Long Short Term Memory)神经网络对F0进行非线性映射，并与源基频的结构信息进行融合的方法。首先通过成对基频F0训练稳定的BLSTM网络，然后通过融合转换后的F0'和保留原始结构信息的F0，得到最终所需基频，最后进行语音合成，进而提高转换语音与目标语音的相似度。同时通过与其他几种线性变换基频的语音转换算法仿真对比，验证了本文所提方法能够在一定程度上降低转换语音的机械音问题，同时能够有效提升转换语音的相似度。

关键词：语音转换，BLSTM神经网络，基频转换，非线性映射

1. 引言

语音转换是指在不改变语言内容的前提下，通过修改源说话人的语音特征参数使其听起来像目标说话人的语音[1]。通常情况下语音转换系统主要由模型训练部分和语音转换部分组成。模型训练阶段，主要是通过平行语料获取源说话人和目标说话人语音特征参数的映射关系。语音转换阶段则是利用训练好的转换模型将源说话人的任意说话内容转换成目标说话人的声音。

近几年随着神经网络和机器学习等算法的兴起，语音转换技术也获得了长足的发展。因为神经网络在处理大量不能用规则或公式描述的原始数据时，表现出了极大的灵活性和自适应性[2]，而语音数据很多时候难以用公式详细描述，所以基于神经网络的语音转换算法备受关注，相关的语音转换的算法也层出不穷。大多数基于神经网络的语音转换基本流程如图1所示。

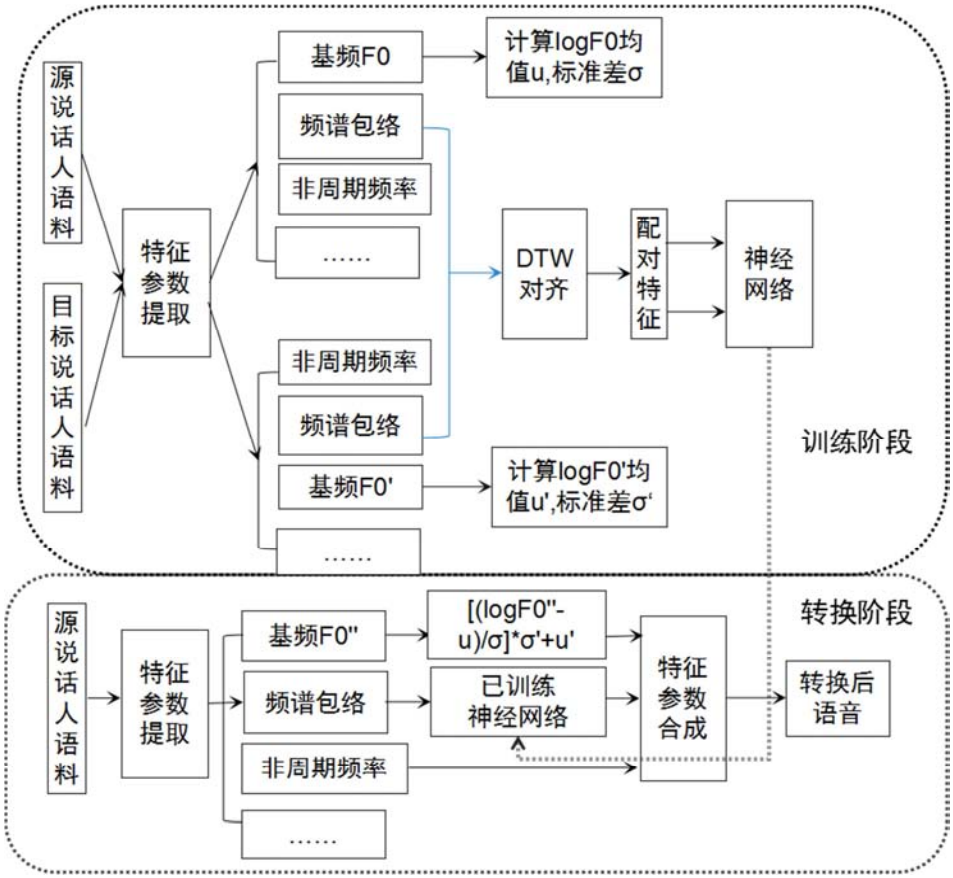


图1 基于神经网络的语音转换基本流程框图。

语音转换主要分为训练和转换两个阶段，在这两个阶段中主要的技术或任务又集中在语音特征参数提取方法和神经网络模型的选择上，这两种技术的不同组合也就产生了不同的语音转换系统。以VCC2016（Voice Conversion Challenge 2016）中众多语音转换算法为例。特征参数提取的工具主要有：STRAIGHT、AHOCODER、HTS等。神经网络模型则包含：DNN、RNN、BLSTM[5]以及HMM等等。

在大多数基于神经网络的语音转换算法中，更多侧重的是对语音频谱参数的提取和变换，而对于基音频率F0的变换则基本采用均值方差的线性变换，例如：文献[3]提出的利用深度BLSTMW网络进行语音转换，其F0采用的线性转换。文献[4]提出了将基频作为wavenet的条件概率进行训练，但是其仍采用的是线性转换后的基频，进行的训练。文献[5]等等也都是用的线性转换基频。虽然语音转换中频谱参数的转换效果极大程度的决定了语音的转换效果，但是基音频率F0作为包含语音韵律特征的参数，其转换后与目标基频的相似度也将影响着语音转换的相似度。而且语音信号的处理过程都是非线性过程，显然线性转换算法在精确性方面存在明显缺陷[2]。而且线性转换基频有时会造成转换语音的“机械音”或突然变调等问题。

因此也有一些学者和专家开展了对基频的非线性转换研究，比如文献[6]中作者提出采用音高目标模型来实现F0转换，通过GMM的方法训练转换模型，但是其并未考虑上下文信息最F0的影响，所以存在一定弊端。文献[7]提出使用DBLSTM-RNN网络转换F0，这个网络能够兼顾上下文信息，但是其原始F0的结构信息为保留，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。还有其他学者提出了F0的其他转换方法，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F0的结构信息，依然得不到理想的语音转换效果。

正是基于上述语音，本文提出了基于 BLSTM 神经网络实现基频 F0 的非线性的融合转换算法，将通过 BLSTM 转换后的基频 F0 与原始基频 F0 中的结构信息相融合，保持其与转换频谱的一致性，进而提高转换语音与目标语音的相似度。

2．基于BLSTM的F0转换融合算法

因为语音转换后基频F0与目标语音的F0越相近，则利用其合成的语音与目标语音相似度越高。通过神经网络训练时极易丢失F0中的结构信息，有效保留源语音基频F0中的结构信息，可以减少因频谱和基频信息不一致而造成的转换语音的质量差的问题。均值方差的线性转换虽然能够较好的保存源语音基频F0的结构信息，但是其转换精度较低，且线性变换难以保证源语音和目标之间的多样性变化。

结合语音转换的基频F0的非线性转换相关研究中，陈芝等人[2]提出利用了RBF网络训练基频，但是其未考虑训练过程中的F0结构信息，所以转换效果一般。文献[8]中提出了针对F0的轨迹转换，采用这种方法重点是对F0的轨迹进行建模而不是F0的本身值，而且在转换过程中采用了插值处理，在精确度上存在一定的不足，同时选取DNN网络结构模型没有考虑语音信号上下文之间的关系信息等。

2.1．BLSTM(双向长短时记忆网络)

考虑到基频F0是语音韵律的一个重要特征，其一定程度上依赖于上下文信息[9]，因此在选取神经网络时，我们采用了BLSTM的神经网络。BLSTM是RNN神经网络的一种改进，RNN能够通过使用循环连接来建模一定量的上下文，并且原则上可以从以前的输入的整个历史映射到每个输出[10]。双向RNN示意图如图2所示。

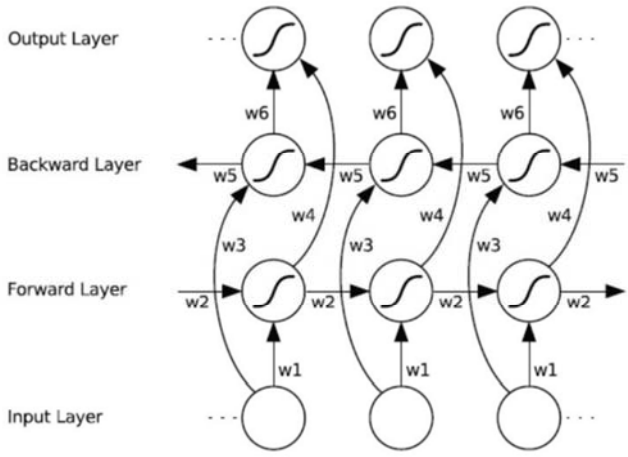


图2 双向循环神经网络在时间上展开[11]。

同时在对传统的递归神经网络中的误差流的分析中发现，远距离上下文传递时，因为反向传播的错误积累或衰减随着时间的推移而爆炸或消失。克服这一问题的一种有效方法是引入长短时存储器体系结构[12]，它能够在较长的时间内在线性存储单元中存储信息，并且可以学习与分类任务相关的最佳上下文信息量。

图3展示了一个长短是记忆模块。文献[12]详细介绍了其组成的原理及其向前和向后的推算公式。

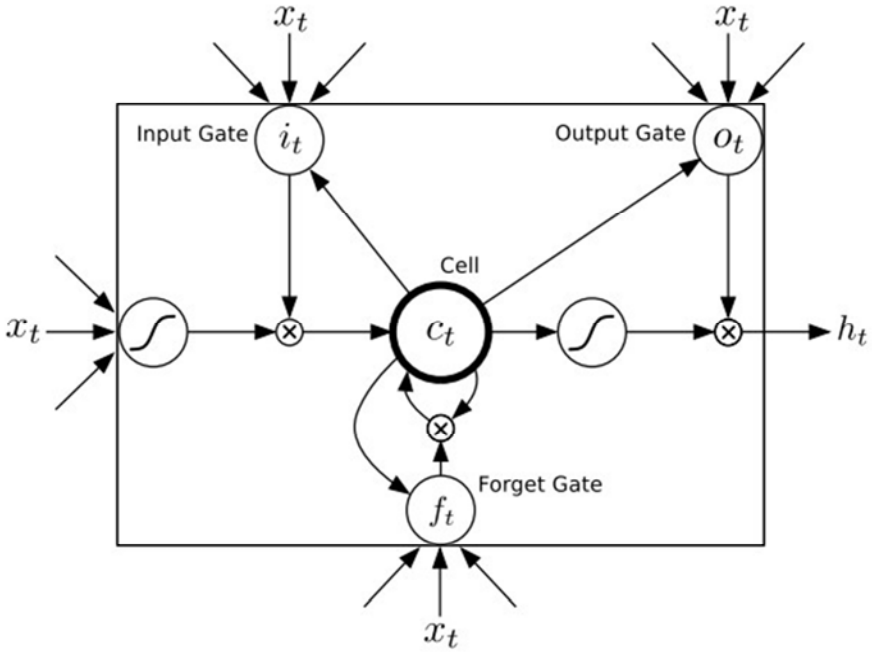


图3 长短是记忆模块。

图4则展示的是一个具有存储单元的LSTM网络体系结构[13-14]。相比于其他神经网络结构模型，BLSTM更能够兼顾序列间的特征，同时保证上下文之间的联系，所以F0转换也选取了该网络模型。

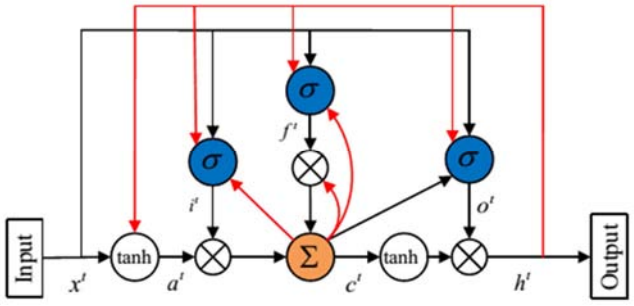


图4 具有存储单元的LSTM网络体系结构[8]。

2.2．F0的非线性转换及融合

文献[3]和文献[15]中采用的是目前基于神经网络的语音转换中常用的F0转换公式如式(1)所示：

$$P_t^{(Y)} = \frac{P_t^{(X)} - u^{(X)}}{\delta^{(X)}} \times \delta^{(Y)} + u^{(Y)} \tag{1}$$

式(1)中，$P_t^{(Y)}$ 和 $P_t^{(X)}$ 分别表示转换后的logF0和原始logF0。$u^{(X)}$ 和 $u^{(Y)}$ 是转换前后的均值，$\delta^{(X)}$ 和 $\delta^{(Y)}$ 是转换前后的标准差。其均值和方差均来自训练数据，通过对大量训练数据的统计和计算所得。

为了克服上述提到的线性转换算法带来的不足，同时也为了更好的兼顾语音转换过程中基频与频谱的对应关系，本文提出了基于BLSTM神经网络的F0融合转换算法，其转换流程如图5所示。将转换融合后最终得到的F_0，代替语音转换过程中线性变换的F0，然后进行语音合成，进而提升语音转换质量。

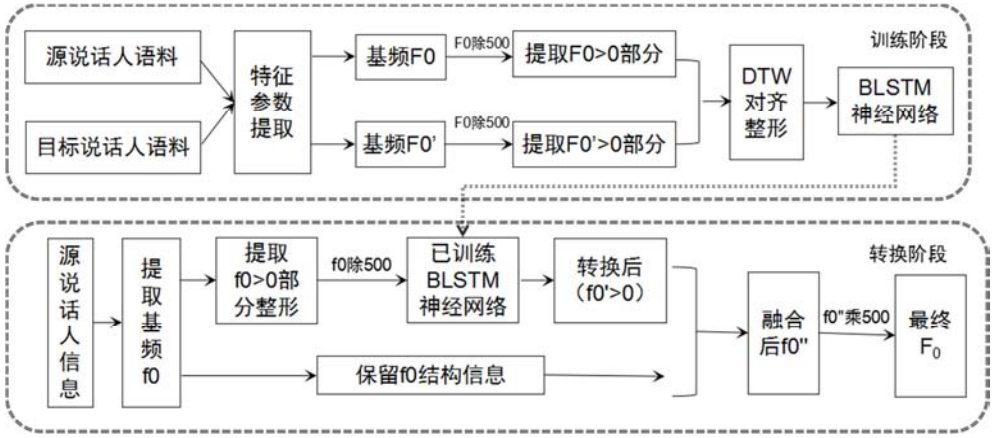


图5 基于BLSTM的F0转换融合流程图。

结合图5流程现对基于BLSTM网络的F0转换步骤作具体说明：

Step 1:利用STRAIGHT工具对源说话人和目标说话人平行语料进行基频F0和F0’的提取

Step 2:对提取到成对的基频进行归一化处理。因为后续BLSTM神经网络训练中选取的激活函数是tanh函数，其作用区间主要是[-1 1]，所以需要对基频进行归一化处理，通常来说,男声的基音频率分布范围为0～200Hz,女声的基音频率分布范围为200～500Hz[16]。所以选取500作为除数，比较合理。

Step 3:分别提取F0和F0’中大于零部分，然后进行动态时间规整DTW。因为即使同样语料内容，不同说话人因说话间隔停顿不同，也不能保证数据完全对齐，而神经网络训练时则要求数据需要一一对应，所以需要进行DTW。但是在训练过程中如果基频中存在过多“0”元素，将会造成训练网络误差偏大，如果不先提取各自大于零部分，直接进行DTW，会是对齐误差较大不利于得到较好的转换网络。

Step 4:DTW后得到对齐的基频相关的数据对，为了能够使网络训练数据丰富和加快收敛，对数据格式进行变换，将一维数据重复，变换成二维数据。然后送入BLSTM神经网络进行训练直至网络收敛。得到稳定的转换网络。

Step 5:提取测试用的源目标说话人语句的基频信息f0。一方面将原始f0进行保存，因为这部分f0包含了其与频谱相对应的结构信息。另一方面同时对f0进行归一化（除以500），提取f0>0部分和整形等预处理，最终将处理后的数据送入训练好的转换模型中得到转换后f0’。

Step6: 将提取的包含结构信息的f0与经过BLSTM网络转换后的f0’融合得到f0’’，然后乘以500得到最终转换F0， 用最终基频与其他网络训练的语音参数进行合成，得到最终的转换语音。

融合过程：

1)提取f0中大于零部分的坐标位置，然后将其坐标依次存入数组A[i]

2)将转换后的f0’的每一个值依次赋给A[i]

3)最后将A[i]按照其坐标位置返回原始f0的位置处，这样就可以得到最后包含结构信息的F_0。

3．实验及结果

为了验证本文所提算法能否提升基频转换精度和改善语音转换质量，本文也分别通过不同的仿真实验进行验证。

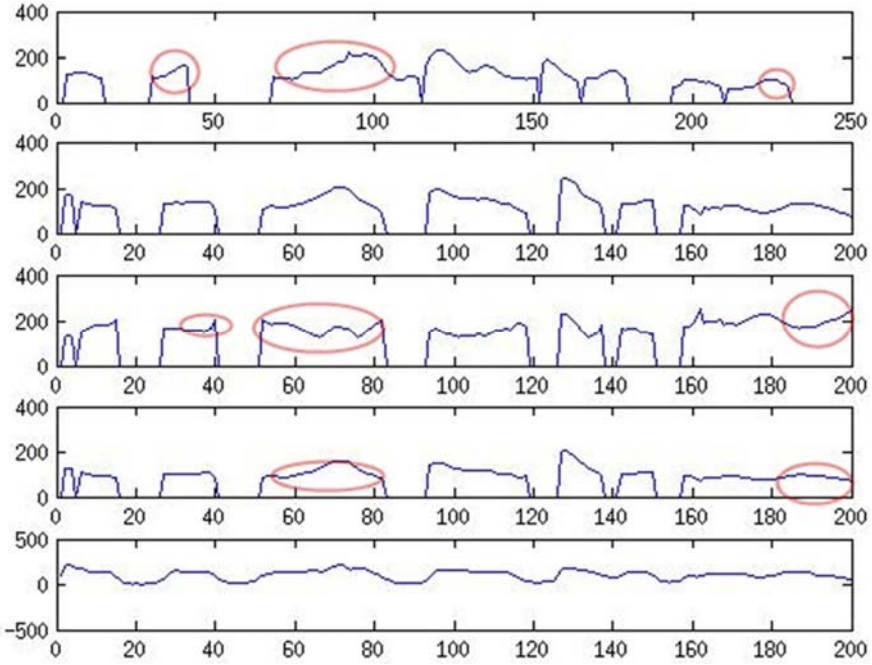
本次实验中选取的数据库是VCC2018官方数据库（我们选取了vcc2018_training文件夹中的部分数据作转换测试）。

实验中用到的软件为matlab2015a，Anaconda3.0, python3.5（本次实验中语音参数的分析和合成部分通过matlab中的STRAIGHT工具实现的，同时对数据的预处理部分也是在matlab中进行的，而BLSTM的神经网络部分则是通过python进行搭建的）,实验是基于Linux14.0操作系统上进行。

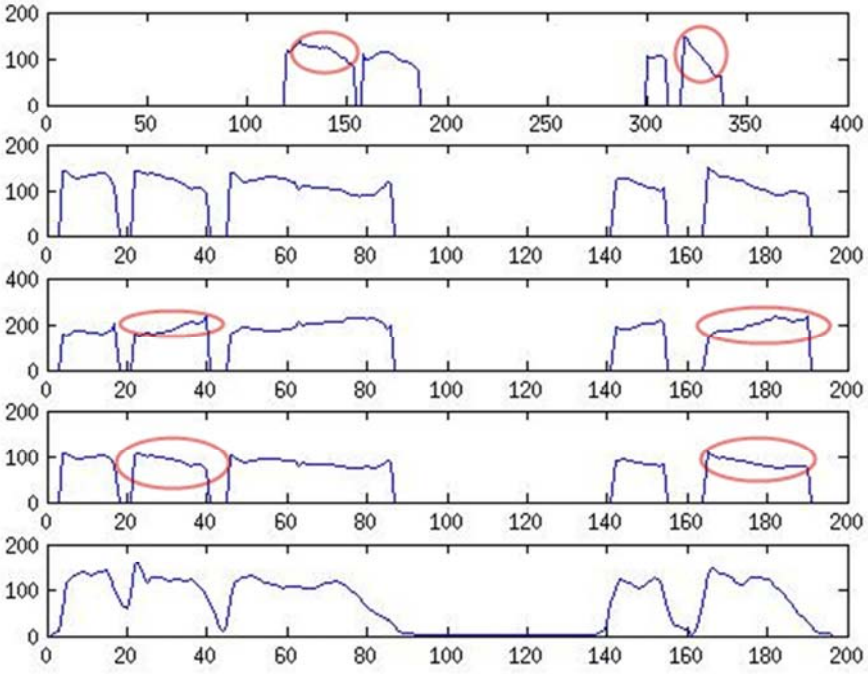
3.1．客观评价指标

为了更加详细和准确的比较转换后基频F0的准确程度，实验中将线性变换转后后的基频和本文所提算法转换基频，以及未采用结构信息直接进行非线性转换基频的结果进行了对比。实验结果如图6所示。

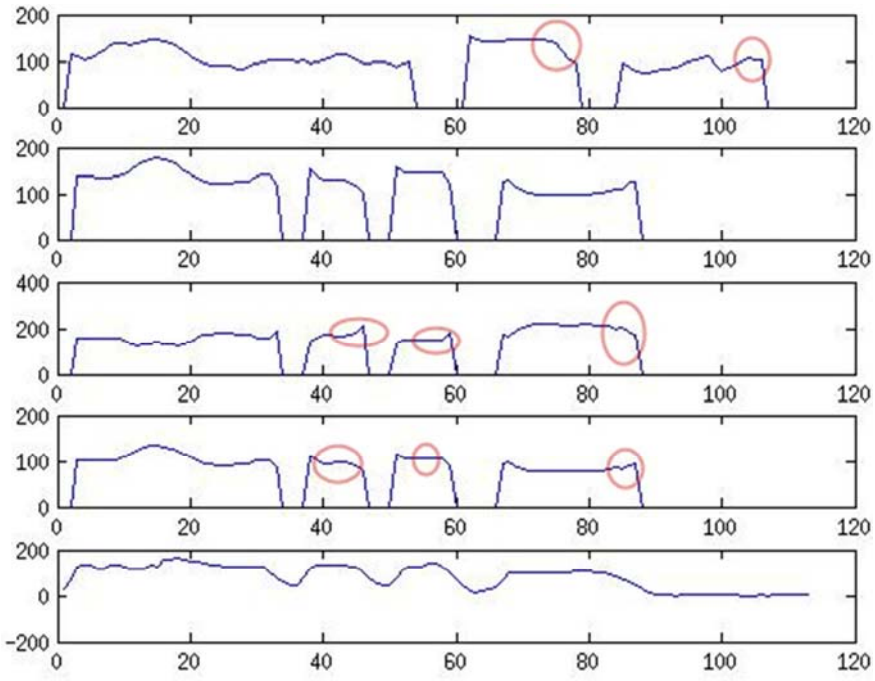
观察图6，其中(a)(b)(c)分别表示随机选取的三句语音的基频对比图。每幅图的第一行表示目标基频，即：要转换成的最终结果。第二行表示源说话人基频，第三行表示采用均值方差线性转换的基频，第四行表示本文所提算法的转换效果，第五行表示直接对F0进行网络训练，忽略F0结构信息的转换结果。



(a)



(b)



(c)

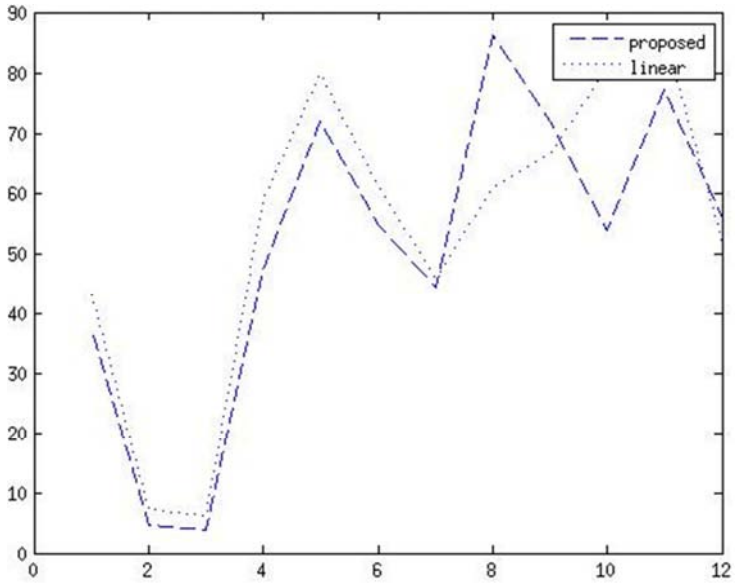
注：图中横坐标单位为帧，纵坐标单位为Hz。

图6 转换基频对比图。

观察图6(a)中红色圆圈部分，其中第一行的目标基频呈现逐渐上升趋势部分，而第三行的线性转换后基频只在尾部骤升，这样就会造成合成语音中出现“怪调”问题，而第四行本文所提算法则没有出现这种问题。同时，图6(a)、图6(b)和图6(c)中红色圆圈部分，都有需要转换后目标基频呈现下降趋势的部分，部分的线性转换反而呈现上升趋势，而需要转换后呈现上升部分线性转换后则又呈现下降趋势，但可以发现本文所提算法则很好地实现了预期的转换效果。

同时，对比观察每幅图最后一行，其表示的为保留原始基频的结构信息实现的BLSTM非线性网络转换，因为没有采取保留原始F0的结构信息，直接将对齐后的基频对数据进行训练，就会在训练和转换阶段造成含零部分的错误映射，在转换后的基频中体现出来就是连续不间断的值，虽然轮廓上保持了基频的大致走向，但其合成语音效果极差，不能够有效用于语音转换。

为了更加准确地展示线性转换和本文转换算法的差距，实验中随机抽取十二组线性转换和本文算法转换的基频，分别计算它们与目标基频的标准差。绘制如图7所示的曲线。



注：图中横坐标表示语句组数，纵坐标表示标准差数值。

图7 误差曲线对比图。

图7表示线性转换F0和本文所提算法转换F0与目标F0的标准差。由图7可以看出本文所提算法实现的F0转换有10组误差小于线性转换误差，有两组误差大于线性转换误差，说明在本文所提的转换算法具有一定的鲁棒性，也进一步证明了所提算法的优越性，对于其中误差较大的两组，其可能原因为：DTW在含零情况下的对齐刚好优于未含零情况下的对齐。

3.2．主观评价指标

在语音转换评价指标中，MOS(Mean Opinion Score)分是一种被广泛认可的评价指标。它是通过多人对转换语音质量的打分来衡量转换语音的效果好坏的指标。

MOS主观评分的分值1-5由受试者主观打分一般把GMM方法的MOS分作为beseline，对于不同的测试可以考虑把GMM的分数对齐从而在不同的MOS测试中统一基准打分基准：

4－5分：优秀(excelent) 很好，听的清楚，延迟很小，交流流畅

3－4分：良好(good) 稍差，听的清楚，延迟小，交流欠缺顺畅，有点杂音

2－3分：一般(fair) 还可以，听不太清，有一定延迟，可以交流

1－2分：差(poor) 勉强，听不太清，延迟较大，交流重复多次1分以下 很差(bad) 极差，听不懂，延迟大，交流不通畅

本文将基于BLSTM、DNN、GMM三种神经网络语音转换算法进行了实现，然后将其中线性转换基频部分替换为本文所提方法，然后再合成转换语音，并将其与线性转换进行了对比。得出关于相似度的MOS分如表1所示（表中所得数据为20组数据打分的平均值）。

表1 转换语音相似度MOS得分表。

	线性转换	本文算法
GMM方法	2.10	2.25
DNN方法	2.32	2.64
BLSTM方法	2.85	3.13

结合表1可以看出，采用本文算法转换F0后的语音相似度都获得主观上的提升。说明本文所提算法对改善语音转换质量起到了一定的作用。(注：文中得分并非最优得分，因为实验中未对网络结构进行优化和调整，但根据控制变量法，只是单纯改变基频F0的转换形式即可提升得分，说明基频F0确实能够改善语音转换质量)

4．结论

本文利用BLSTM的神经网络对语音转换中的基频F0进行非线性处理，通过对基频F0的前期预处理和后期转换后的融合处理，极大程度上提升了转换F0与目标F0的相似度，重点对原始F0的结构信息进行了保留，使得F0通过神经网络的非线性映射能够较好的实现。同时验证了F0的改善能够提升转换语音与目标语音的相似度。

致谢

本文为国家自然科学基金项目(61471394)的阶段性成果之一。

参考文献

[1] Xiaohai Tian, Zhizheng Wu, S. W. Lee, and Eng Siong Chng. Correlation-based Frequency Warping for Voice Conversion [C] International Symposium on Chinese Spoken Language Processing. 2014:211-215

[2] 陈芝,张玲华.基频轨迹转换算法及在语音转换系统中的应用研究[J] 南京邮电大学学报(自然科学版),2010,10,30(5):83-87

[3] L Sun, S Kang, K Li et.. al. Voice conversion using deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks [C]. IEEE International Conference on Acoustics, 2015:4869-4873

[4] Jumpei Niwa, Takenori Yoshimura, Kei Hashimoto. et.. al. Statistical Voice Conversion based on WaveNet [C]. Speech and Signal Processing (ICASSP) 2018:5289-5293

[5] 王民，杨秀峰，要趁红.基于PSO优化GRNN的语音转换方法[J].计算机工程与科学.2018,4(40):752-756

[6] Y Kang, J Tao, B Xu. Applying Pitch Target Model to Convert F0 Contour for Expressive Mandarin Speech Synthesis [C]. IEEE International Conference on Acoustics, 2006, 1:I-I

[7] Huaiping Ming1, Dongyan Huang1, Lei Xie. et.. al. Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion[C]. Interspeech, 2016:2453-2457

[8] Hy Quy Nguyen · Siu Wa Lee · Xiaohai Tian. et.. cl. High quality voice conversion using prosodic and high-resolution spectral features [J]. Multimedia Tools & Applications, 2016, 75 (9): 5265-5285

[9] Meyer, G. A.: The Semantics of Stress and Pitch in English. The Faculty Association, Utah State University (1961)

[10] Martin Wollmer, Angeliki Metallinou, Nassos Katsamanis et.. al. Analyzing the memory of BLSTM Neural Networks for enhanced emotion classification in dyadic spoken interactions [C]. IEEE International Conference on Acoustics, 2012, 1 (15):4157-4160

[11] James Zhang's Blog. 双向长短时记忆循环神经网络详解（Bi-directional LSTM RNN）[DB/OL]https://blog.csdn.net/jojozhangju/article/details/51982254

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory.[J] Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] Xiangang Li and Xihong Wu. Improving long short-termmemory networks using maxout units for large vocabulary speech recognition. [C] Speech and Signal Processing (ICASSP), in Acoustics, 2015 IEEE International Conference on. IEEE, 2015, pp. 4600–4604.

[14] Zhiying Huang, Jian Tang, Shaofei Xue et.. al. Speaker adaptation OF RNN-BLSTM for speech recognition based on speaker code.[C]IEEE International Conference on Acoustics, 2016:5305-5309

[15] 解伟超.语音转换中声道谱参数和基频变换算法的研究[D].南京邮电大学.2013.04

[16] 张超琼,苗夺谦,岳晓冬.基于语音基频的性别识别方法及其改进[J].中文科技论文在线.http://www.paper.edu.cn