第52卷第5期
2020年10月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 52 No. 5
Oct. 2020

DOI:10.16356/j.1005-2615.2020.05.014

一种基于特征融合的耳语音向正常音的转换方法

庞聪,连海伦,周　健,王华彬,陶　亮

(安徽大学计算智能与信号处理教育部重点实验室,合肥,230039)

摘要: 使用耳语音的频谱包络来预估正常音的基频特征,这类算法在对正常音基频预测的准确性上存在一定不足,在合成语音自然度方面存在着明显欠缺,有时会出现音调失常等问题。本文提出一种声学特征融合的方法,通过双向长短期记忆(Bi-long short-term memory, BLSTM)深度网络来逐帧预测正常音基频。首先,使用STRAIGHT模型和相关代码,分别对耳语音和正常音语料进行预处理,提取耳语音的梅尔倒谱系数(Mel-scale frequency cepstral coefficient,MFCC)、韵律及谱包络特征,正常音的基频与谱包络特征。然后使用BLSTM深度网络,分别建立耳语音和正常音谱包络特征之间映射关系,以及耳语音MFCC、韵律及谱包络特征对正常音基频F_0的映射关系。最后根据耳语音的MFCC、韵律及谱包络特征获得对应的正常音基频和谱包络,使用STRAIGHT模型合成正常音。实验结果表明,相较于仅使用谱包络估计基频,采用此种方法引入语音韵律和MFCC的融合特征是对基频特征的良好补充,解决了音调失常的现象,转换后的语音在韵律上更加接近正常发音。

关键词: 语音转换;特征融合;韵律模型;STRAIGHT模型;双向长短期记忆

中图分类号: TN912.3　　**文献标志码:** A　　**文章编号:** 1005-2615(2020)05-0777-06

Method for Transforming Whisper to Normal Speech with Feature Fusion

PANG Cong, *LIAN Hailun*, *ZHOU Jian*, *WANG Huabin*, *TAO Liang*

(Key Laboratory of Computational Intelligence and Signal Processing, Ministry of Education, Anhui University, Hefei, 230039, China)

Abstract: Currently, in reconstruction of normal speech from whispered speech based on neural network, the spectral envelope of the whisper is often used to estimate F_0 characteristics of the normal speech. Such algorithms have certain deficiencies in the accuracy of F_0. There is a clear lack of naturalness, and sometimes the pitch distortion occurs. This paper proposes a method for predicting the F_0 of normal speech frame by frame using the Bi-long short-term memory(BLSTM)deep network with the acoustic fusion feature of normal speech. Firstly, the STRAIGHT model and related codes are used to preprocess the whisper and the normal speech corpus. Respectively, extract the Mel-scale frequency cepstral coefficient (MFCC), rhythm and spectral envelope of the whisper speech and the F_0 and spectral envelope of the normal speech. Secondly, the BLSTM deep network is used to establish a mapping relationship between spectrums of whisper and normal speech, and a mapping relationship between MFCC, rhythm and spectral envelope features of whisper speech and F_0 of normal speech. Finally, according to MFCC, rhythm and spectral envelope features of whisper

基金项目: 国家自然科学基金(61301295)资助项目;安徽省自然科学基金(1708085MF151)资助项目;安徽高校自然科学基金(KJ2018A0018)资助项目;安徽大学科研训练计划(J10118520444)资助项目。

收稿日期: 2019-06-06;**修订日期:** 2020-01-05

通信作者: 周健,男,副教授, E-mail: jzhou@ahu.edu.cn。

引用格式: 庞聪,连海伦,周健,等. 一种基于特征融合的耳语音向正常音转换的方法[J]. 南京航空航天大学学报,2020,52(5):777-782. PANG Cong, LIAN Hailun, ZHOU Jian, et al. Method for transforming whisper to normal speech with feature fusion[J]. Journal of Nanjing University of Aeronautics & Astronautics,2020, 52(5):777-782.

speech, the F_0 and spectral envelope of the corresponding normal speech are obtained, and the normal speech is synthesized using the STRAIGHT model. The experimental results show that compared with the estimation of the F_0 using only the spectral envelope, the introduction of fusion features of phonetic rhythm and MFCC is a good complement to the F_0 features, which solves the phenomenon of pitch disorders and the converted speech is closer to normal speech in rhythm.

Key words: voice conversion; feature fusion; prosodic model; STRAIGHT model; bi-long short-term memory

耳语音是一种有别于正常语音的常见发音方式,发声时声带不振动,肺部气流通过半开的狭窄声门产生类似噪声的气声[1]。耳语音转换就是将耳语音转换为正常音的技术,在很多领域有着广阔的发展前景。现有的耳语音转换技术主要分为基于规则的转换和基于统计模型的转换。近几年,随着深度学习和神经网络算法的崛起,语音转换技术也获得了较大的发展。神经网络在处理大规模的杂乱无规则原始数据时,能有效地提取高阶特征,而表现出极大的自适应性。有研究表明,语音信号是一种典型的稀疏信号,因此耳语音和正常音的频谱具有稀疏性。考虑到耳语音和正常音的不同发音特性,耳语音和正常音特征之间存在非线性映射关系,所以基于神经网络的语音转换算法被广泛采用。

传统耳语音向正常音的转换算法主要是通过提取耳语音的谱包络特征、正常音的基频与谱包络特征,利用神经网络建立耳语音和正常音低维谱包络特征之间的映射关系,以及耳语音低维谱包络特征和正常音基频之间的映射关系[2-4]。然而,基于神经网络的语音转换方法更多关注频谱包络的变换,并不重视语音的韵律特征。文献[5-6]所训练的深度双向长短期记忆(Bi-long short-term memory, BLSTM)神经网络方法只完成了语音频谱包络的转换映射,再与线性变换后的基频F_0特征相结合,最后利用STRAIGHT语音合成算法合成正常音。在耳语音转换中,虽然频谱参数变换的效果对最终耳语音转换的效果影响较大,但耳语音的韵律特征也同样影响着转换后正常音的基频特性、语速特性和能量特性[7]。然而,现有算法在对正常音基频的预测转换的准确性上并不十分理想,转换后语音的自然度和可懂度有待进一步提高。

基频F_0是语音韵律的一个重要特征[8],虽然耳语音不存在基频F_0,但人耳可以通过幅值包络等韵律特征来辨别音调[9]。本文提取从听觉感知角度描述短时梅尔倒谱系数(Mel-scale frequency cepstral coefficients, MFCC)特征以及基频相关、能量相关和时长相关的韵律特征。这些韵律特征表征的是声门信息,与表征听觉特征的MFCC有着良好的互补性。本文利用MFCC和韵律的融合特征与谱包络结合来估计正常音的基频,然后利用从耳语音谱包络估计的正常音谱包络合成正常音。实验结果表明使用这种方法估计正常音基频F_0的准确性有所提高,合成语音在音高,音调等方面更接近自然发音。

1 双向长短时记忆神经网络

语音信号是典型的时序信号,语音基频和频谱包络的预测和转换在一定程度上依赖前后相邻帧的语音特征。图1给出了一个最基本的双向循环神经网络(Bi-recurrent neural network, Bi-RNN)的结构图。

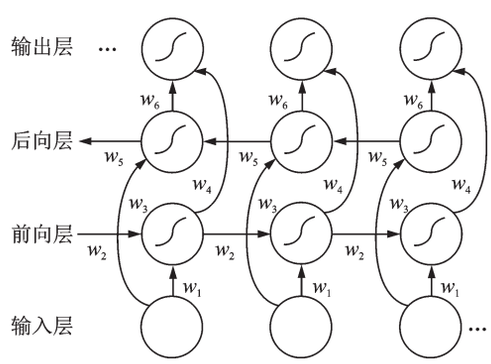

图1 Bi-RNN结构图

Fig.1 Structural diagram of Bi-RNN

在图1中可以看出,Forward层和Backward层共同连接着输出层,其中包含6个共享权值w_1~w_6。在每个时刻结合Forward层和Backward层的相应时刻输出结果得到最终的输出,即有

$$h_t = f(w_1 x_t + w_2 h_{t-1}) \tag{1}$$

$$h'_t = f(w_3 x_t + w_5 h'_{t+1}) \tag{2}$$

$$o_t = g(w_4 h_t + w_6 h'_t) \tag{3}$$

由于传统RNN存在无法解决长程依赖问题的致命缺陷,在远距离上下文传递中,结果经过多次传播后梯度趋向于消失,训练好的神经网络对语音转换效果较差。为解决这个问题,本文采用双向长

短时记忆神经网络，引入LSTM记忆模块：1个记忆单元，用于存储网络时序状态；输入门、遗忘门和输出门用于实现信息的保护和控制，以及1个LSTM记忆块（图2）。

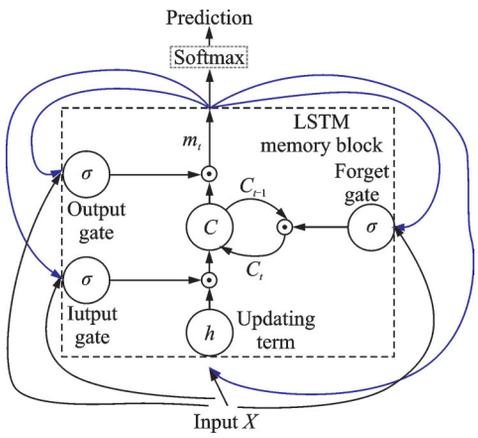

图2　LSTM记忆块

Fig.2　Memory block of LSTM

2　MFCC、韵律以及谱包络特征提取

2.1　MFCC特征提取

MFCC是在Mel标度频率域提取出的倒谱参数，Mel标度描述了人耳频率的非线性特性，它与频率的关系可用下式近似表示

$$\mathrm{Mel}(f)=2\,595\times\lg\left(1+\frac{f}{700}\right) \tag{4}$$

语音特征参数MFCC提取过程如图3所示。图3中，语音经过以下几个阶段处理得到MFCC参数[10]：

(1)语音信号的预处理，主要有预加重、分帧、加窗、对加窗后的各帧信号进行快速傅里叶变换得到各帧的频谱，并对语音信号的频谱取模平方得到语音信号的功率谱。

(2)将能量谱通过一组Mel尺度的三角形滤波器组，Mel三角带通滤波器组频率响应定义为

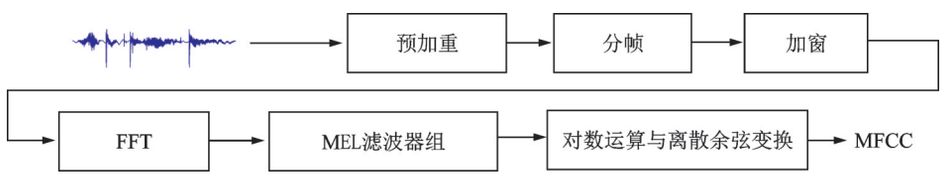

图3　MFCC参数提取过程

Fig.3　Extracting process of MFCC

$$H_m(k)=\begin{cases}0 & k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1)(f(m)-f(m-1))} & f(m-1)\leqslant k\leqslant f(m)\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1)(f(m)-f(m-1))} & f(m)\leqslant k\leqslant f(m+1)\\ 0 & k\geqslant f(m+1)\end{cases} \tag{5}$$

式中：$\sum_{m=0}^{M-1}H_m(k)=1$。

(3)计算每个滤波器组输出的对数能量

$$s(m)=\ln\left(\sum_{k=0}^{N-1}|X_a(k)|^2H_m(k)\right)\quad 0\leqslant m\leqslant M \tag{6}$$

(4)经离散余弦变换得到MFCC系数

$$C(n)=\sum_{m=0}^{N-1}s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right)\quad n=1,2,\cdots,L \tag{7}$$

式中：L为MFCC系数阶数，通常取12～16；M为三角滤波器个数。

2.2　韵律特征提取

韵律特征是语音在音高、音强等方面所表现出来的抑扬顿挫的特性。

本文将共振峰、能量相关特征及短时平均过零率作为韵律特征输入来提高对正常音的基频轨迹预测的准确性，提高合成语音的自然度和可懂度。

对于能量相关特征，文中提取了语音的短时能量轨迹，短时能量变化率及短时平均振幅[11]。

2.3　语音谱包络提取

本文使用STRAIGHT模型来提取耳语音的频谱包络以及正常音的基频F_0和频谱包络。

设$x(t)$为原始语音信号，对其进行分帧、加窗等预处理后，经过短时傅里叶变换得到语音的短时谱$X(\omega,t)$。STRAIGHT工具可对语音短时谱进行频域的自适应内插平滑，使用补偿窗和三角窗去除语音短时谱在时域和频域上的周期性[12]，获得每个语音帧的频谱包络为

$$\overline{X}(\omega,t)=\sqrt{g^{-1}\left(\iint_D h_t(\lambda,\tau)g(|X(\omega-\lambda,t-\tau)|^2)\mathrm{d}\lambda\mathrm{d}\tau\right)} \tag{8}$$

式中：τ_0为基音周期；f_0为基音频率。$h_t(\lambda,\tau)=\frac{1}{4}\left(1-\left|\frac{\lambda}{\omega_0(t)}\right|\right)\left(1-\left|\frac{\lambda}{\tau_0(t)}\right|\right)$，$\omega_0(t)=2\pi f_0(t)$，

$-\omega_0(t)\leqslant\lambda\leqslant\omega_0(t),-\tau_0(t)\leqslant\lambda\leqslant\tau_0(t)$。

3 基于声学融合特征的耳语音到正常音转换

3.1 基于BLSTM的耳语音到正常音特征映射

本文使用两个深度BLSTM网络实现耳语音到正常音的特征映射。

图4中,左上部分是从耳语音谱包络到正常音谱包络的深度神经网络BLSTM1。网络将耳语音和正常音的频谱包络分别作为输入和输出,隐藏层深度为4,网络结构为[128-256-256-128]。

左下部分是从耳语音谱包络、MFCC、韵律和频谱包络的融合特征到正常音基频F_0的神经网络BLSTM2。该网络隐含层结构类似BLSTM1,输入层为耳语音语料的333维数据,包括257维谱包络特征,70维MFCC特征以及6维韵律特征(包括1维短时平均能量、1维短时平均过零率、1维短时平均振幅和3维共振峰信息)。

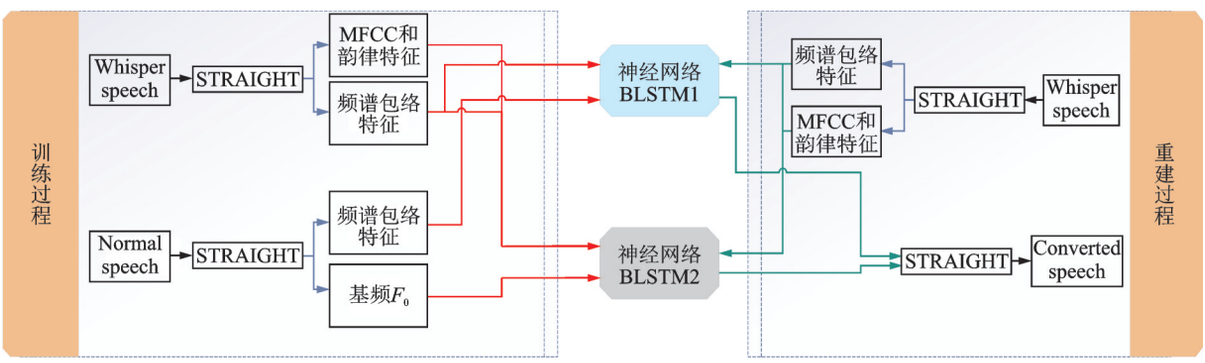

图4 语音转换框架图

Fig.4 Diagram of voice conversion framework

3.2 基于STRAIGHT模型重建正常音

对于输入的耳语音,先对其进行预加重等预处理,使用STRAIGHT工具获得耳语音每帧的频谱包络序列$\{\overline{X_1^{\omega}},\overline{X_2^{\omega}},\cdots,\overline{X_L^{\omega}}\}$,再使用韵律特征提取工具提取耳语音的每帧MFCC和韵律特征,总共76维。将谱包络送到图4训练好的网络BLSTM1,将所有特征融合处理后,送入图4训练好的深度网络BLSTM2中,获得转换后的正常音逐帧谱包络序列$\{\overline{X_1^{\hat{s}}},\overline{X_2^{\hat{s}}},\cdots,\overline{X_L^{\hat{s}}}\}$和逐帧基频序列$\{F_{01},F_{02},\cdots,F_{0L}\}$,之后通过STRAIGHT工具可获得合成后的正常音。

3.3 实验结果及分析

实验选择TIMIT语音库中的348句耳语音及对应的正常音为实验数据。采样频率F_s为8 kHz,16位PCM存储,声学特征提取时帧长取40 ms,帧移5 ms,FFT的点数为512,由此得到的谱包络特征为257维,耳语音的MFCC和韵律融合特征为76维。为了验证本文提出的特征融合方法在正常音基频预测方面的效果,实验采用基频误差分析和客观评价指标来评价本文方法性能。

本次实验的语音参数提取、数据预处理以及语音参数合成部分通过MATLAB中的STRAIGHT工具箱实现,而深度BLSTM网络的训练和转换部分是在Tensorflow框架上搭建,基于Ubuntu16.04操作系统进行。

3.3.1 基频误差分析

由于耳语音没有基频,并且谱包络信息包含较少的基频信息,因此在使用耳语音谱包络估计正常音基频时会造成部分基频缺失。为解决这一问题,本文提出声学特征融合的方法,引入MFCC与韵律信息进行补充。

图5(a)—(c)分别表示随机选取的一句语句的基频对比图。BLSTM_SP表示仅使用谱包络估计基频的方法在下文中用,BLSTM_SP$^+$表示本文提出的基于声学特征融合的耳语音预测正常音基频方法。

图5中圆圈部分显示:使用BLSTM_SP方法转换的基频曲线缺失大量基频特征,这样就会造成合成语音中出现音调失常问题,而BLSTM_SP$^+$方法中则没有出现这种问题,实验结果表明基于声学特征融合来估计正常音基频的方法是更加有效的。

为了更加准确地展现线性转换和本文转换算法的差距,实验中随机抽取了一句语句,分别使用BLSTM_SP方法与本文提出的BLSTM_SP$^+$方法转换的基频与谱包络合成正常音,绘制如图6所示语音语谱图。

观察图6中的圆圈部分,在相同谱包络的情况下,如果基频缺失,语谱图就会出现声纹模糊甚至声纹缺失。

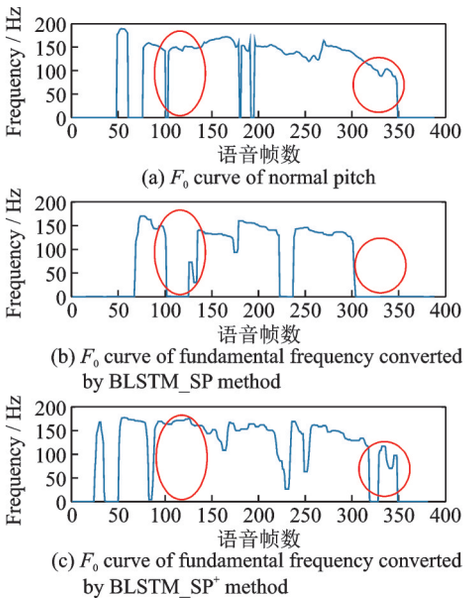

图 5　基频曲线对比图

Fig.5　Comparison of fundamental frequency curves F_0

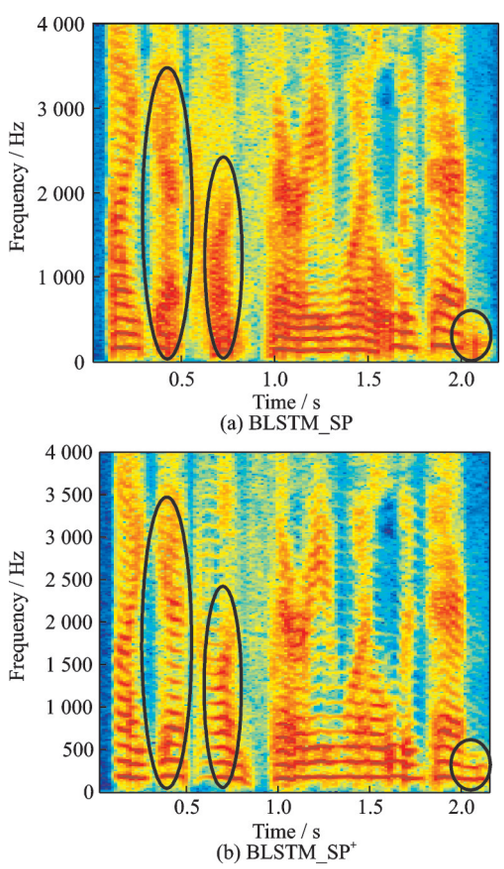

图 6　语谱图

Fig.6　Spectrogram

3.3.2　客观评价指标

客观评价采用短时可懂度测量指标(Short-time objective intelligibility，STOI)[13]和倒谱失真度(Cepstral distortion，CD)[14]以及客观语音质量评估(Perceptual evaluation of speech quality，PESQ)。

STOI是对语音可懂度的一种客观评估方法，在STOI算法中，输入为正常音信号$x(n)$和转换后的语音信号$y(n)$，输出结果为一标量值d，值越大，意味着重建后的语音可懂性越高[15]。倒谱失真度CD是一种常见的频谱转换客观评价方法，公式如下

$$\mathrm{CD}=\frac{10}{\log 10}\sqrt{2\sum_{d=1}^{D}\left(C_d-C_d'\right)^2} \tag{9}$$

式中：C_d和C_d'分别为转换语音和参考语音的第d维梅尔倒谱系数。D代表梅尔倒谱维度，实验中设置为26。用各帧的平均值作为该段语音的CD值，CD值越大，表明转换语音与参考语音之间差异越大。

PESQ是一种客观语音质量评估算法。在PESQ算法中，同时输入正常音信号$x(n)$和经过转换的语音信号$y(n)$，算法输出的取值范围为0～5，值越大，意味着转换后的语音质量越高。表1给出了实验得到的各项客观评价指标。

表 1　客观评价指标表

Table 1　Objective evaluation indicator

Method	CD	STOI	PESQ
BLSTM_SP	5.014 1	0.576 1	1.182 3
BLSTM_SP$^+$	4.924 1	0.577 4	1.212 5

表1中实验结果显示，本文提出的BLSTM_SP$^+$方法效果较原方法在CD，STOI及PESQ指标上都有所改善，转换后语音质量提高。

4　结　　论

实验通过在常规耳语音谱包络估计正常音基频算法的基础上引入耳语音韵律特征和MFCC特征，将其融合特征作为BLSTM深度网络输入来估计正常音基频F_0。实验结果发现，基于声学特征融合的方法补充了谱包络缺失的基频特征，改善了基频缺失现象，提高了正常音基频预测准确性。这种方法为正常音预测研究提供了新的思路，如何进一步提高正常音基频的预测准确度，使合成的语音更贴近正常发音以满足实际应用的需求仍然是后期需要深入研究的问题。

参考文献：

[1]　SHA Jun，CHEN Xueqin，YU Yibiao. Comparison of performance between normal and whispered speech in Chinese isolated word recognition[C]//Proceedings of International Conference on Signal Processing.[S.l.]：IEEE，2015：545-548.

[2]　周健，窦云峰，刘荣敏，等.采用低维特征映射的耳语

音向正常音转换[J].声学学报,2018,43(5):855-863.
ZHOU Jian, DOU Yunfeng, LIU Rongmin, et al. Whisper to normal conversion based on low dimension feature mapping[J]. Acta Acustica, 2018, 43(5): 855-863.

[3] 王民,苏利博,王稚慧,等.采用STRAIGHT模型和深度信念网络的语音转换方法[J].计算机工程与科学,2016,38(9):1950-1954.
WANG Min, SU Libo, WANG Zhihui, et al. Voice conversion using STRAIGHT model and deep belief networks[J]. Computer Engineering & Science, 2016,38(9): 1950-1954.

[4] AHANGAR M, GHORBANDOOST M, SHATMA S, et al. Voice conversion based on a mixture density network[C]//Proceedings of Applications of Signal Processing to Audio & Acoustics. [S.l.]: IEEE, 2017: 329-333.

[5] SUN L, KANG S, LI K, et al. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks[C]//Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing.[S.l.]: IEEE, 2015: 4869-4873.

[6] LIAN Hailun, HU Yuting, ZHOU Jian, et al. Whisper to normal speech based on deep neural networks with MCC and F_0 features[C]//Proceedings of 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP). [S.l.]: IEEE, 2018: 1-5.

[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] LUO Z, TAKIGUCHI T, ARIKI Y. Emotional voice conversion using deep neural networks with MCC and F_0 features[C]//Proceedings of 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). [S.l.]: IEEE, 2016: 1-5.

[9] 沙丹青,栗学丽,徐柏龄.耳语音声调特征的研究[J].电声技术,2003(11):4-7.
SHA Danqing, LI Xueli, XU Boling. Study on the characteristics of the tones in whispered Chinese[J]. Audio Engineering, 2003(11): 4-7.

[10] SITHARA A, ABRAHAM T, DOMINIC M. Study of MFCC and IHC feature extraction methods with probabilistic acoustic models for speaker biometric applications[J]. Procedia Computer Science, 2018,143: 267-276.

[11] 刘翠.语音信号韵律特征提取及其应用研究[D].江门:五邑大学,2014.
LIU Cui. Speech signal prosodic features extraction and its application research[D]. Jiangmen: Wuyi University, 2014.

[12] KAWAHARA H. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited[C]//Proceedings of 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. [S. l.]: IEEE, 1997: 1303-1306.

[13] TAAL C H, HENDRIKES R C, HEUSDENS R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech[C]//Proceedings of IEEE International Conference on Acoustics Speech & Signal Processing. [S.l.]: IEEE, 2010: 4214-4217.

[14] PHAM T D, BYUNGSUB S. A cepstral distortion measure for protein comparison and identification [C]//Proceedings of International Conference on Machine Learning & Cybernetics. [S.l.]: IEEE, 2005: 4214-4217.

[15] 彭晓腾.语音可懂度客观评价策略的研究[D].呼和浩特:内蒙古大学,2016.
PENG Xiaoteng. The research on objective strategies of speech intelligibility[D]. Hohhot: Inner Mongolia University, 2016.

(编辑:孙静)