文章编号：2096-1618(2019)01-0001-06

基于基音跟踪的语音增强研究

蔡 良， 夏秀渝， 陆 雄， 孙文慧
(四川大学电子信息学院，四川 成都 610065)

摘要：在移动通信、语音识别、基于语音的语音交互等领域，采集的语音信号往往混杂具有谐波结构的噪声，因此语音增强都有非常重要的应用价值。语音的能量大部分集中在浊音段，浊音具有谐波结构。基于实际混合声音在时频域具有近似稀疏性特点，提出一种基于基音跟踪的语音增强算法，利用基音特征尽可能地恢复语音的谐波结构同时抑制噪声信号能量来达到提升语音信噪比的目的。首先对混合声音流进行切分、浊音段提取，接着对浊音段信号进行多基频提取，并利用维特比解码找出主导基频，使用BP神经网络对主导基频进行是否人声基频的判别，最后利用梳齿滤波器重构浊音段语音或抑制干扰音。仿真实验表明，算法能够从混有音乐和背景噪声的混合音频中提取语音，语音信噪比增益平均达8 dB。

关 键 词：语音增强；维特比算法；基音跟踪；多基频提取
中图分类号：TN912.35 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2019.01.001

0 引言

随着人工智能技术不断发展[1]，基于语音的人机交互技术让人们与计算机之间的交互变得更加高效，设备在采集语音的过程中总是不可避免的混入各种噪声干扰，使机器对语音处理效果急速下降，所以语音增强技术在工业上有着非常重要的意义。目前应用最广泛的增强算法有谱减法、维纳滤波法，两种算法去除平稳噪声效果不错，但噪声非平稳时去噪性能急剧下降。使用语音盲分离技术可以从平稳或非平稳噪声中分离出语音，但盲分离系统为多传声器系统，盲分离算法复杂，系统成本高。计算听觉场景分析(CASA)是一种模拟人耳听觉特性，实现混合声音自动分离的方法，目前主要的计算听觉场景分析方法都是基于基音线索，但低信噪比情况下，主导基频的计算容易出错，另外算法计算量很大，计算速度上还不能满足实时应用的要求。

实际的声学环境非常复杂，除一般的环境噪声外，干扰音(如音乐、鸟鸣等)往往和语音的频率分布范围重叠，也是随机非平稳的。混合声音，(如音乐、语音等)信号大多具有谐波结构，所以实际混合声音在时频域往往具有近似稀疏性[2]。基于浊音是语音的主要成分，具有基音及其谐波结构，文中提出一种基于主导基音跟踪和判别的语音增强算法，算法首先对混合语音进行切分、浊音段提取，之后利用PEFAC[3]多基频提取算法提取候选基频；再利用维特比解码算法进行主导基频提取；然后利用BP神经网络进行人声基频识别，最后重构语音以提升语音质量。论文通过一系列仿真实验验证算法的有效性。

收稿日期：2018-05-08

1 算法原理

文中的语音增强算法主要针对具有谐波结构的浊音信号，通过多基音提取和跟踪技术，人声和非人声基音的判别，利用谐波重构的方法从混合声音中分离出语音。算法框图如图1所示。

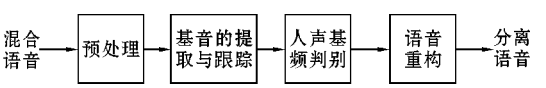

图1 算法原理框图

1.1 预处理

预处理部分包括预加重、分帧、FFT、切分等。语音是短时平稳的，预处理部分主要做了有声无声的判断、切分和清浊判断等工作，其具体框图如图2所示。

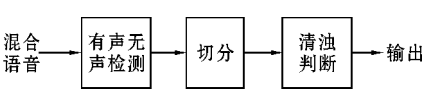

图2 预处理框图

预处理部分中的预加重主要是为了去除发声过程中口唇辐射的影响，提升语音高频部分。端点检测可将音频流中的无声段去除，只保留有声段，以减少后续处理的计算量。文中使用短时能量作为端点检测的特征参数，进行端点检测[4]。如下式：

$$\begin{cases}\text{有声段} & \text{if} \quad E_i > T \\ \text{无声段} & \text{if} \quad E_i \leqslant T\end{cases} \tag{1}$$

其中 E_i 为第 i 帧语音短时能量,统计了多段语音有声段和无声段的能量,与手工标记的有声和无声段进行比对,在这里选取 $T=0.7*\sum_{i=1}^{f_n}E_i/f_n$ 的效果最好。声音是由一系列声音事件组成,音频切分的目的是将音频流切分成一系列相对独立的短时音频段落,各音频段落间听觉谱差异较大,段内谱差异较小。对于具有谐波谱结构的语音浊音或音乐信号来说,段间频谱的变化和基频的变化是对应的。文中采用 *DIS* 度量距离算法[5]来切分语音,*DIS* 距离是一种综合语音段段间均值和方差的度量方法,来表征语音段落之间的差距:

$$DIS=\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{tr(\sum_1)+tr(\sum_2)} \tag{2}$$

其中,$\boldsymbol{\mu}_1$,$\boldsymbol{\mu}_2$ 代表前后两段语音特征的均值矢量,$tr(\sum_1)$,$tr(\sum_2)$ 表示不同前后两段语音特征的标准差。音频特征参数选用短时幅度谱。通过逐帧滑动计算,可以得到 *DIS* 曲线,提取曲线上极大值点(如果极大值相隔太近,只取其中一个)。若极大值点的 *DIS* 值大于预设门限,那么判断为分割点,否则舍去。*DIS* 分割之后,语音就被切分为各个相对独立的短时音频段落,这些音频段内的基音变化范围较小。

有声段包含清音和浊音,语音只有浊音段才有基频特征,因此在有声段内还要进行清浊判断。浊音段信号自相关函数具有一定的周期性,而清音不是,因此利用每帧音频信号自相关函数一定时滞取值范围内的峰值大小与一个门限比较来进行清浊判断:

$$\begin{cases}\text{浊音段} & \text{if} \quad a > T_V \\ \text{清音段} & \text{if} \quad a \leqslant T_V\end{cases} \tag{3}$$

其中 a 为自相关函数基频范围内第一个峰值大小,通过统计第一个峰值和该帧能量的大小,在比例为 0.2的时候效果最佳,因此这里取 $T_V=0.2E_i$。通过以上处理,将音频流自动切分为静音段、有声段,有声段进一步切分成小段,段内还切分成清音浊音。

1.2 基音提取与跟踪

语音的浊音、音乐等信号具有特定的基频谐波结构,混合音频信号某帧往往存在多个基频,采用改进的 PEFAC[6] 多基音提取算法提取基音,然后用维特比解码算法跟踪主导基音轨迹。PEFAC 算法是一种能在低信噪比条件下提取多基频的有效算法。一帧基频为 f_0 的浊音段,信号在频域上可以表示为

$$Y(f)=\sum_{k=1}^{K}a_k\delta(f-kf_0) \tag{4}$$

其中 a_k 为 k 次谐波上的系数。将其转换到对数频域上可以表示为

$$Y(q)=\sum_{k=1}^{K}a_k\delta(q-\log k-\log f_0) \tag{5}$$

其中 $q=\log f$,谐波间的间隔由 $\log k$ 来决定。在对数域上设计一个梳齿滤波器为

$$h(q)=\sum_{k=1}^{K}\delta(q-\log k) \tag{6}$$

然后做卷积运算,令 $A(q)=h(q)\times Y(q)$,则 $A(q)$(基频显著度函数)在 $q=\log f_0$ 处会产生一个峰值。上式中的 $h(q)$ 是一个理想梳齿滤波器,实际上由于分帧加窗处理,会使得信号谐波会有一定的宽度,因此实际采用的 $h(q)$ 为

$$h(q)=\int\log(\gamma-\cos(2\pi e^q))\,dq-\log(\gamma-\cos(2\pi e^q))\,dq \tag{7}$$

其中 $\log(0.5)<q<\log(K+0.5)$,K 为峰值的个数,γ 是用来调节滤波器的宽度,在这里取 $\gamma=1.25$。

通过以上处理可以得到每帧信号的基频显著度函数,然后查找峰值得到多个候选基频。为消除基频半频倍频的影响,采用了两步查找法。第一步找出基频显著度函数上最大的峰值点,作为第一个候选基频;第二步在去除显著度函数上第一候选基频倍频和半频位置峰值的基础上,找出第二高峰,作为第二个候选基频。文中每帧提取两个候选基频。

基音跟踪采用维特比算法[7],引入3 个量:基音的观察概率、转移概率和初始概率。基音观察概率由每帧候选基频的显著度计算得到,定义为

$$p(f_m\mid A_i)=\frac{a_{i,s}}{\sum_s a_{i,s}} \tag{8}$$

其中 $a_{i,s}$是第 i 帧的第 s 个候选基频的显著度,A_i 是第 i 帧的显著度函数。基音转移概率 $p(\Delta f)$ 通过统计人声数据集的基频变化率得到,基频变化率定义为

$$\Delta f=\frac{f_i-f_{i-1}}{f_{i-1}} \tag{9}$$

其中 f_i 是当前帧的基频,f_{i-1}为前一帧基频。统计人声数据集中所有浊音段 Δf 的概率密度并归一化得到 $p(\Delta f)$。针对提取基频时可能出现丢失某帧目标基频的情况,将小于0.0001 的概率值取值统一设置为 0.0001。

初始概率 p_i 为每一浊音段内第一帧的观察概率。基音跟踪在各浊音段内进行,找出一条最优的基音序列,使其满足:

$$\hat{S}=\underset{(f_1,f_2,\cdots,f_T)}{\operatorname{argmax}}\left\{\sum_{i=2}^{T}\lg(p(f_m\mid A_i))+\sum_{i=2}^{T}\lg(p(\Delta f))\right\} \tag{10}$$

这样的最优基频序列,就是这一段音频的主导基频轨迹。

1.3　基于 BP 神经网络的人声基频识别

提取的主导基频不一定都是人声基频,因此还要判断每段跟踪的主导基频轨迹是否为人声基频轨迹,为后续采用不同的语音增强方式做准备。人声和非人声的发声机理不同,所具有的音色也不同,可以依据音色区分不同的声音。音色是声音的基本属性,主要由谐波的多寡及其相对强度决定。信号谱包络特征可以较好反映声音音色信息,即声音基音和各次谐音的相对强度决定音色信息,而 MFCC(梅尔倒谱系数,实际应用中最常用的语音特征)就是一种准确描述声音谱包络的一种特征,因此基于 MFCC 参数和合适的分类器可以进行人声和非人声的判别。用于模式识别的分类器有很多,文中采用 BP 神经网络进行[8]人声和非人声识别。

当然不可能直接利用混合声音的 MFCC 进行识别,考虑到浊音、音乐等信号具有谐波结构,混合声音具有近似的稀疏性,因此可以基于目标声源的基音信息利用梳齿滤波器提取与该基频对应的单声源谐波频谱,然后再提取 MFCC 送入 BP 神经网络识别,从而判断该基频是否为人声基频。

具体实现步骤为:

(1)根据每帧的主导基频 f_0,构造一个梳齿滤波器 $H(k)$。

(2)滤波,提取主导基频对应声源的谐波谱:

$$Y(k)=H(k)X(k) \tag{11}$$

其中 $Y(k)$ 为提取的单声源谐波谱,$H(k)$ 为和基频对应的梳齿滤波器,$X(k)$ 为原始的混合语音。

(3)利用提取的谐波谱 $Y(k)$ 计算 12 维的 MFCC 特征参数[4]。

(4)将 MFCC 参数送入 BP 神经网络得到的判别结果。

(5)统计各帧判别结果确定每段音频主导基音轨迹是否为人声基音轨迹。根据1.1节分的相对独立的音频段,在每个音频段内进行主导基频人声判别结果统计。若那一段内,主导基频人声帧数大于 65%,则判定主导基频为人声基频,否则为非人声基频。

BP 神经网络使用分为训练阶段和识别阶段。训练阶段利用纯净语音和干扰音按照上述步骤(1)-(3),提取 MFCC 参数输入神经网络进行训练。识别阶段对每帧混合信号利用上述步骤提取出的 MFCC 参数,送入训练好的神经网络进行人声非人声判别。

1.4　语音重构

声音流分为静音段,清音段和浊音段。利用静音段估计噪声谱,可以采用谱减法去除平稳的环境噪声。语音和干扰音都有清浊音之分,混合语音具有近似的时频稀疏性,因此我们分几种情况分离增强语音。

通过预处理可得到静音段和有声段的位置,利用静音段的音频数据估计噪声幅度谱

$$\hat{N}(f)=\frac{1}{N}\sum_{i=1}^{N}|X_i(f)| \tag{12}$$

其中,$X_i(f)$ 为第 i 帧的语音频谱,N 为静音段的帧数。静音段采用谱减法进行语音增强

$$|\hat{S}_i(f)|=\begin{cases}|X_i(f)|-\hat{N}(f) & \text{if } |X_i(f)|>\hat{N}(\mathrm{f})\\ 0 & \text{if } |X_i(f)|\leqslant\hat{N}(f)\end{cases} \tag{13}$$

$\hat{S}_i(f)$ 为增强后的语音频谱。混合语音中的清音部分可能为语音清音、干扰音清音、混合清音,采用以下方法处理。首先进行是否人声判断,对清音段提取 MFCC 后,送入神经网络判别是否为人声清音。如果是人声清音,可以利用式(13)处理。如果为非人声清音,可以对频谱进行衰减

$$\hat{S}_i(f)=\alpha X_i(f) \tag{14}$$

其中 α 为衰减因子。对于浊音段,提出一种利用梳齿滤波器提取语音谐波的方法,可以近似还原出语音浊音段频谱。由1.2节知,浊音段的频谱主要由语音谐波和干扰音谐波组成[9]

$$X(f)=\sum_{k=0}^{K}a_kS(f-kf_0)+\sum_{k=0}^{K}b_kN(f-kf_1) \tag{15}$$

其中 $S(f)$ 为语音谐波,$N(f)$ 为干扰音谐波。a_k,b_k 为语音和噪声谐波系数。

设计一个对应基频 f_0 的梳齿滤波器 $H(f)$

$$H(f)=h(f)\times r(f) \tag{16}$$

其中 $h(f)$ 为 4 KHz 范围内,对应基频 f_0 的冲击序列

$$h(f)=\sum_{k=0}^{K}\delta(f-kf_0),\ K\in\left[\frac{4000}{f_0}-1,\frac{4000}{f_0}\right] \tag{17}$$

$r(f)$ 为脉冲基本波形

$$r(f)=\begin{cases}0.54-0.46\mathrm{COS}[F/(n-1)],\ 0<f<N\\ 0.01\qquad\qquad\text{,其他}\end{cases} \tag{18}$$

可以调整 $f(f)$ 宽度 N 以适应信号的谐波宽度。

根据人声基频识别,可以识别某段主导基频是人声还是非人声的基音轨迹,对于人声段,要保留谐波结构,用梳齿滤波器($H_i(f)$)滤除干扰谐波。对于非人声段,利用梳齿凹口滤波器($1-H_i(f)$),去除谐波

$$|\hat{S}_i(f)|=\begin{cases}|X_i(f)|H_i(f) & \text{if}\quad \text{人声主导}\\ |X_i(f)|(1-H_i(f)) & \text{if}\quad \text{非人声主导}\end{cases} \tag{19}$$

其中,$|X_i(f)|$ 为混合语音幅度谱,$\hat{S}(f)$ 为增强

之后语音频谱,$H_i(f)$为对应基频梳齿滤波器,$1-H_i(f)$为对应基频的梳齿凹口滤波器。以上算法操作简单,能很好还原语音幅度谱。最后可以结合混合语音相位和增强后语音的幅度谱进行逆傅里叶变换,还原出语音信号。

2 实验结果及分析

为验证语音增强方法的可行性,进行了一系列仿真实验。文中采用乐器音频模拟干扰音,用高斯白噪声模拟环境噪声,模拟场景为具有背景音乐和环境噪声的两人对话场景。使用的人声数据集是从广播电台,一些脱口秀节目等截取的较为纯净的语音,收集了20个不同人的声音,其中男女比例约为1∶1,总时长约为11个小时,由于语音信号主要在低频部分,为了减少计算量,音频降采样至8 KHz,对人声部分利用1.2节的方法提取基频,人工加注基频标签。非人声部分采用 MIR-1K[10] 数据集中音乐伴奏部分,降采样至8 KHz,带有基频标签。引入信干比(SIR)、信噪比(SNR)分别描述干扰音和环境噪声大小:

$$SIR=10\cdot\lg\left(\frac{\sum S^2(t)}{\sum I^2(t)}\right) \tag{20}$$

$$SNR=10\cdot\lg\left(\frac{\sum S^2(t)}{\sum N^2(t)}\right) \tag{21}$$

其中,$I(t)$是干扰声,$S(t)$是语音信号,$N(t)$是环境噪声。

2.1 音频切分及主导基频跟踪实验

在音频库中随机选取了一段纯净男声,中文为“蓝天白云,碧绿的大海”,添加一段干扰声(琵琶独奏),信干比为-10 dB,同时叠加信噪比为20 dB的高斯白噪声模拟环境噪声。按照1.2节所述方法进行音频切分、主导基频跟踪的效果如图(3)所示。

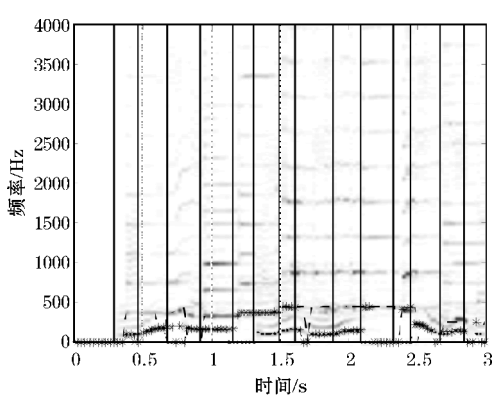

图3 主导基频提取

其中星型线表示提取的主导基频,虚线表示干扰音基频,点状线表示纯净人声基频,背景为混合语音的语谱图,从图可以看出切分之后整个音频被切分成了短时的音节或音符段落,每段音频内部基频变化不是很明显。在低信噪比情况下,主导基频基本分布在人声或干扰音基频上,并能去除倍频和半频的影响。

2.2 人声非人声基频判别实验

为了训练人声基频判别模型,将人声数据集和非人声数据集均分成训练集和测试集,比例约为1∶1。为保证训练和识别数据的一致性,训练时根据语音或音乐的基频标签提取对应音频谐波频谱,然后再提取MFCC特征作为BP神经网络的输入特征。BP神经网络设置如下:采用3层网络结构,输入端为12维MFCC,输出端2个节点,目标误差为e^{-4},学习速率为0.05,迭代次数5000次。为测试模型的性能,将测试集中的语音根据基频标签提取MFCC,送入神经网络中测试以综合识别率β作为评价指标

$$\beta=\frac{N_t}{N} \tag{22}$$

其中,N_t是识别正确的帧数,N为总帧数。实验表明综合识别率达到了82.10%。

2.3 语音增强实验

为了衡量目标语音的增强效果,引入信噪比增益指标[11]

$$SNRG=10\cdot\lg\left(\frac{\sum S^2(f)}{\sum[\hat{S}(f)-S(f)]^2}\right)-10\cdot\lg\left(\frac{\sum S^2(f)}{\sum[X(f)-S(f)]^2}\right) \tag{23}$$

其中$|S(f)|$为纯净语音幅度谱,$|\hat{S}(f)|$为增强之后的语音幅度谱,$|X(f)|$为混合语音幅度谱。

从音频库中随机选取一段男声为“蓝天白云,碧绿的大海”,干扰音为一段琵琶声,高斯白噪声模拟环境噪声,$SIR=-5$ dB,$SNR=20$ dB。利用算法进行语音增强的效果如图(4)、(5)、(6)所示。

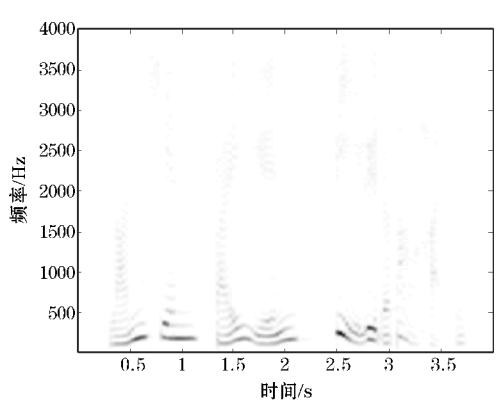

图4 纯净语音语谱图

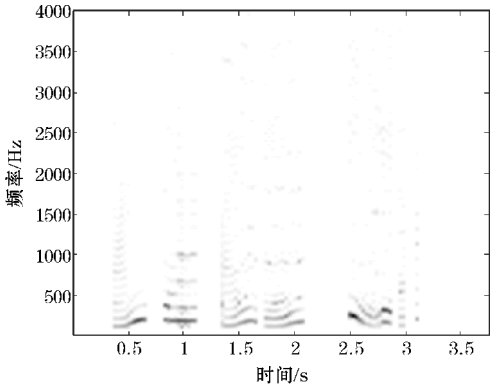

图 5　混合语音语谱图

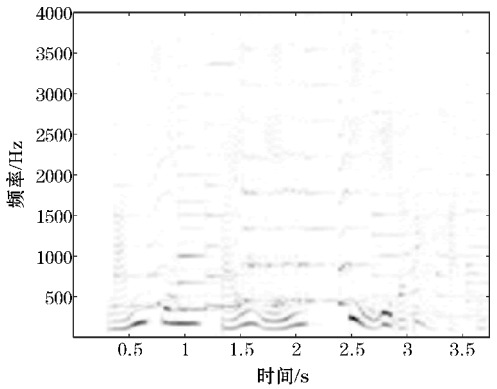

图 6　语音增强后语谱图

从图中可以看出,加入琵琶声后混合语音语谱图上有明显的“条纹状”,处理之后基本还原出语音的语谱图。

在数据集中随机选取 20 段语音,加入不同的乐器声作为干扰音进行试验,同时叠加 $SNR=20$ dB的高斯白噪声,实验结果如表 1 所示。

表 1　不同干扰音情况下语音增强效果对比　　dB

干扰音种类	SIR		
	-5	0	5
琵琶音	11.63	9.75	7.88
钢琴音	10.25	8.37	6.56
笛子音	12.16	9.98	8.02
小提琴	9.59	7.57	5.12
鸡尾酒噪声	8.13	5.57	4.59
平均 *SNRG*		8.344	

从表 1 可以看出,算法能够从混有音乐和背景噪声的混合声音中提取语音,有效提高语音信噪比增益。从表 1 还可看出,算法语音增强的效果和混合语音信干比有关,信干比越低,增强效果越好,信干比越高,增强效果会减弱。增强算法相对于传统的 CASA 增强算法[12],增加了主导基频人声识别,声音增强更具目的性,不再简单地无选择地增强主导音频。仿真实验表明算法平均信噪比增益到达了8 dB左右,相比于传统 CASA 增强算法,提升了1.13 dB,在信噪比较低的情况下,提升比较明显。

3　结论

针对具有谐波结构的语音信号,提出了一种基于基音跟踪的语音增强算法。算法的关键在于如何正确找出语音信号的主导基频。利用 PEFAC 算法对混合语音进行多基频提取,利用基频显著度函数筛选候选基频,并通过维特比解码算法跟踪出主导基音轨迹,结合神经网络判断主导基音是否为人声基频,最后针对混合语音可能的几种混合情况采用不同的处理方法重构出语音,完成语音增强过程。实验结果表明算法可以很好地去除带有谐波结构的背景干扰音,在信噪比很低的情况下,信噪比提升约为8 dB。

参考文献:

[1]　王红. 低信噪比场景下语音增强算法的研究[D]. 合肥:安徽大学,2017.

[2]　胡定禹,郁文贤,江文斌. 基于谐波重建的语音增强算法的研究[J]. 信息技术,2017(11).

[3]　Gonzalez S,Brookes M. Pefac-a pitch estimation algorithm robust to high levels of noise[J]. IEEE/ACM Transactions on Audio Speech & Language Processing,2014,22(2),518-530.

[4]　宋知用. MATLAB 在语音信号分析与合成中的应用[M]. 北京:北京航空航天大学出版社,2013.

[5]　孙彦楠,夏秀渝. 基于深度神经网络的关键词识别系统[J]. 计算机系统应用,2018,27(5):41-48.

[6]　Gonzalez S, Brookes, M. PEFAC-A Pitch Estimation Algorithm Robust to High Levels of Noise [J]. IEEE Press,2014.

[7]　韩纪庆. 音频信息检索理论与技术[M]. 北京:科学出版社,2011.

[8]　吕菲,夏秀渝. 基于方位特征的听觉选择性注意计算模型研究[J]. 自动化学报,2017,43(4):634-644.

[9]　胡定禹,郁文贤,江文斌. 基于谐波重建的语音增强算法的研究[J]. 信息技术,2017,(11).

[10]　Chao-Ling Hsu,Prof. Jyh-Shing Roger Jang. MIR-1KDataset[OL]. http://sites. google. com/site/unvoicedsoundseperation/mir-1k,2009.

[11]　夏秀渝,何培宇. 基于声源方位信息和非线性时频掩蔽的语音盲提取算法[J]. 声学学报,

2013(2):224-230.
[12] 王雨,林家骏,袁文浩,等. 基于改进基音跟踪算法的单通道语音分离[J]. 华东理工大学学报(自然科学版),2013,39(3):338-344.

Research on Speech Enhancement based on Pitch Tracking

CAI Liang, XIA Xiuyu, LU Xiong, SUN Wenhui
(College of Electronic and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: In the fields of mobile communication, speech recognition and voice-based voice interaction, etc., the collected speech signals are often mixed with noise with harmonic structure, so speech enhancement has very important application value. Most of the speech energy is concentrated in the voiced segment, and the voiced speech has a harmonic structure. Based on the fact that the actual mixed-sound shows approximate sparse characteristics in time-frequency domain, this paper proposes a speech enhancement algorithm based on pitch tracking, which use the pitch feature to restore the harmonic structure of the speech as much as possible while suppressing the noise signal energy to achieve the purpose of improving the speech signal to noise ratio. Firstly, the mixed sound stream issegmented and the voiced segmentis extracted. Then, the multi-pitch extraction is performed on the voiced segment signal. The dominant pitch is found through Viterbi decoding, and the BP neural network is used to discriminate whether the dominant pitch is vocal pitch. Lastly, The comb-tooth filter is used to reconstruct the speech in the voiced segment or to suppress the interference. The experimental results showed that the algorithm successes extracting speech from mixed-audio which is mixed with music and background noise, and the ratio of speech signal to noise gains 8dB in average.

Keywords: speech enhancement; viterbi algorithm; pitch tracking; multi-pitch extraction