第 57 卷 第 7 期 2008 年 7 月
1000-3290/2008/57(07)/4244-07

物 理 学 报
ACTA PHYSICA SINICA

Vol.57, No.7, July, 2008

基于人耳听觉模型的自动嗓音评估方法*

王 迪[1)2)†] 付 强[2)] 杨 琳[2)] 于 萍[3)] 颜永红[2)] 冯 稷[1)]

1)(中国科学院物理研究所,北京 100190)

2)(中国科学院声学研究所中科信利语音实验室,北京 100080)

3)(中国人民解放军总医院,北京 100853)

(2007 年 7 月 15 日收到;2007 年 12 月 4 日收到修改稿)

在嗓音评估系统的长元音谐噪比分析中,针对传统方法在普通傅里叶变换域上进行谐波成分计算并且需要对样本进行人工选择切分的情况,提出了一种新谐噪比计算方法,能够自动切分出长元音中稳定部分,并采用了更贴近人耳听觉模型的时频分析办法,使对长元音的分析能够更稳定更贴近人耳主观听觉.同时由于没有人工干预,使得评估标准更加统一,结果更加客观.

关键词:嗓音评估,听觉模型,长元音分析,谐噪比计算

PACC:4300,4370,8736,8734

1. 引 言

嗓音疾病逐渐成为现代人常见多发的疾病之一,嗓音健康也越来越受到人们的重视.病变嗓音的检测及评估对嗓音疾病的治疗有着至关重要的意义.声带病变会使声带的振动模式发生变化,从而引起人嗓音的异常,检测这种变化就可以判断声带的疾病种类以及病变程度.但人的发声过程是一个复杂的物理学及生理学过程,受到诸多因素的影响,其机理至今还有很多不明确的地方,因此如何进行病变嗓音的检测与评估一直是个复杂的课题.

传统的侵入式检测办法采用喉镜等医疗设备深入患者喉部进行检测,这样不仅会使患者产生不良反应,而且无法检测到声门的振动情况和整个发声过程,从而无法检测出某些非器官性病变.相对而言,基于声学信号分析的评估办法采用非接触式测量,能够记录患者自由发声完整过程,并分析声门在不同的发音模式下声门振动的异常情况,因此相对具有很大优势.

基于声学信号分析的评估方法分为主观评估和客观评估两种类型.主观评估采用嗓音医学专家对嗓音样本进行听觉打分,然后根据得分判断声带的病变种类及程度.这类方法的优点是可以利用专家的经验进行判断,比较符合人的听觉模式.但是也存在很多局限性,比如为了保证评估的有效需要对同一样本进行不同时间的多次评估来降低评估过程中主观性的影响,这样不仅费时费力,而且其受主观影响过大,标准难以统一,结果很难复现.

客观评估方法利用数字信号处理技术,将信号在不同的变换域上进行分解,提取出能反映嗓音病变程度的参数进行分析判断.这种方法能够对样本进行快速而客观的分析,评判标准统一,评判过程和结果可以复现,从而大大拓展了嗓音评估的实际应用范围.

由于稳定性高以及受干扰较小,评估系统一般选择长元音进行分析.相应的医学研究表明[1—4],病变的声带会在病人发浊音时产生一些非周期性的振动,而且这种非周期性随着病变程度的严重而愈发明显.因此分析的主要评判准则是发音周期性的稳定程度.在反映这种稳定性的参数中,最常采用的参数有基频微扰(jitter,反映基频的稳定性)、幅度微扰(shimmer,反映幅度的稳定性)、谐噪比(HNR,反映谐波与噪声成分的比值)等参数.研究表明这些参数中谐噪比数值的大小与声带病变程度的匹配度最高[1,2],这是由于病变造成的声音嘶哑度很大程度上

* 国家重点基础研究发展规划(973)项目(批准号:2004CB318106),国家自然科学基金(批准号:10574140,60535030),国家高技术研究发展计划(863 计划)(批准号:2006AA010102,2006AA01Z195)资助的课题.

† 通讯联系人. E-mail:dwang@hccl.ioa.ac.cn

取决于声音中的噪声和谐波成分的比值.因此谐噪比能否计算准确直接影响到了对患者嗓音评估的准确度.

传统的谐噪比一般在时域上利用周期信号的自相关性,估计出信号的基频周期.然后根据基频周期的位置,将基频周期附近相关性强的部分作为谐波成分,而将相关性较弱或不相关的部分作为噪声来计算谐噪比[1].这种方法存在着一定缺陷:首先,对于一些较为沙哑或病变程度严重的嗓音样本来说,基频周期的估计不仅容易出现偏差,有时甚至无法得出.这样就无法计算出有效的谐噪比数值[5].其次,这些谐波计算是在普通时频域上进行的,这与真实的人耳感知存在很大的差别,因此最后计算结果与嗓音医学专家的打分进行匹配度分析并不合理.

为此,本文提出了一种基于人耳听觉模型的自动嗓音评估方法.此方法利用了人的听觉特性,在听觉谱上对长元音信号进行相关性分析.然后依据听觉场景(CASA)分析中的听觉流的概念,将样本中的谐波看成是不同的听觉流成分加以分离提取最后计算谐噪比.由于此种谐噪比估计方式充分考虑到人耳的听觉特性,因此能更加有效地与主观听觉打分联系起来.而且对于一些难以估计出基频位置的样本,也能进行有效分析.

2. 听觉感知分析和特征提取

2.1. 传统的谐噪比计算方法

设被分析信号 $x(t)$ 为短时平稳信号,则相应的自相关系数 $r_x(\tau)$ 为

$$r_x(\tau) = \int x(t)x(t+\tau)\mathrm{d}t, \tag{1}$$

式中 τ 为时延.$r_x(\tau)$ 在 $\tau = 0$ 时有着全局最大值.若除 $\tau = 0$ 外,当 $\tau = nT_0$ 时还存在着一些局部的极值,则认为此信号存在周期性,其中 T_0 即为周期.

设 r_{H} 和 r_{N} 分别为谐波相关性和噪声相关性,由于谐波成分的相关性和噪声的不相关性(类似白噪声),当 $\tau = 0$ 时有

$$r_x(0) = r_{\mathrm{H}}(0) + r_{\mathrm{N}}(0). \tag{2}$$

而当 $\tau = T_0$ 时有

$$r_x(T_0) = r_{\mathrm{H}}(T_0) = r_{\mathrm{H}}(0), \tag{3}$$

则谐噪比(HNR)为

$$\mathrm{HNR} = 10 \times \log_{10}\left(\frac{r_{\mathrm{H}}(0)}{r_x(0) - r_{\mathrm{H}}(0)}\right). \tag{4}$$

为了对不同响度的声音分析尺度相同,计算前一般会先将自相关系数归一化,然后再进行计算.

在周期性明显的嗓音样本(如图 1 所示)中,采用此种方法可以得出有效结果.但随着声门病变程度的加剧,样本受到的噪声干扰越来越大,信号的周期性变得非常不明显(如图 2 所示).在这样的情况下,这种方法无法估计出基频周期,因而不能对样本进行谐噪比分析.

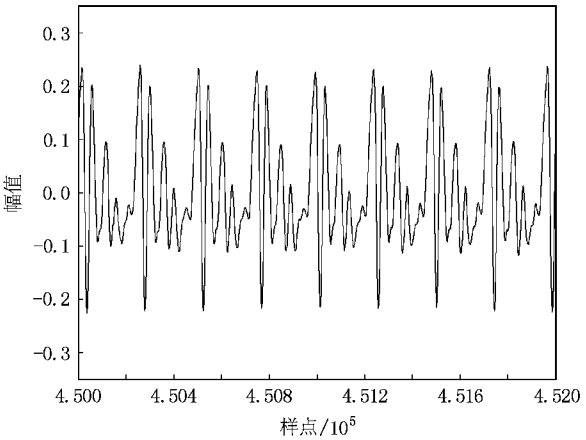

图 1　典型正常长元音样本

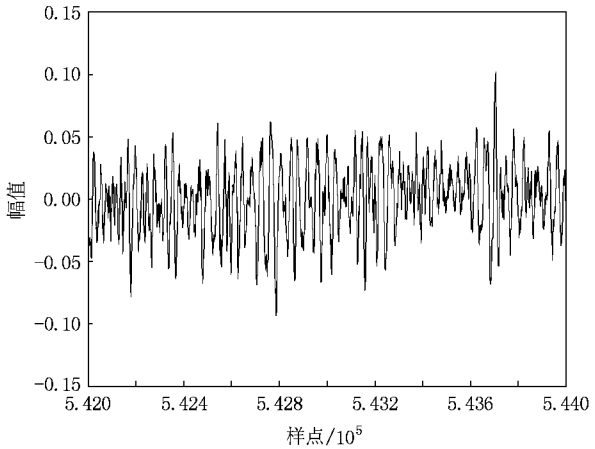

图 2　典型病变长元音样本

后来又有一些人陆续提出一些新的计算方法[6−9],比如利用反复迭代的方法,或者采用倒频谱来估计 HNR 等等.这些方法大都依赖于基频位置的检测,且并未能考虑到人耳听觉谱上对信号的感知.

2.2. 听觉谱域上的谐噪比计算方法

在计算机听觉场景分析(CASA)[10,11] 对音频流的分析中认为,对于混合音频中若干声源来说,同一

声源所发出的声音应该有连续性.这种连续性在时频域上即表现为音频成份在相邻时间上与相邻频带上的连续性.Meddis 等[12]就曾经尝试利用这种相关性,运用听觉模型将信号分解为多个频带,然后计算各个频带内的自相关系数来分离混叠语音信号的基频.

本文将这种分析方法应用在病变嗓音样本的分析中.一般研究表明[1,5],随着嗓音病变程度的增加,样本中类似于白噪的成分将逐渐增强,而谐波的成分将逐渐减弱.由于谐波的成分相关性较强,而白噪声成分相关性较弱,因此利用听觉场景分析中对音源的相关性分析可以有效提取样本中的谐波.

为了分离音频中的谐波和噪声成分,先将信号在听觉谱域上进行分解,然后利用时间域上的自相关系数和通道之间的交叉互相关系数计算出一个系数,用来表示时频域分解下每个部分与周围部分的相关程度.再设定一个阈值,将大于阈值的成分归为谐波成分,否则归为噪声成分.

2.2.1. 听觉谱域上的信号分解

为了在听觉谱域上提出信号的谐波成分,首先利用听觉模型对信号进行分解.经过分解后的信号在耳蜗谱域上由若干时频单元(time-frequency unit)组成,其中每个单元即代表某一帧的某个频带.

1)分帧 数据按照 20 ms 进行分帧,帧间有 50% 即 10 ms 的叠接.这种分帧方法兼顾了语音音素的平稳性和连续性,在语音分析中较为常见.

2)中耳和外耳模型:中耳和外耳对声音信号在 1.5—5.0 kHz 范围内有 10—20 dB 的提升,可以利用预加重方式来大致模拟其压力增益,设原始信号 $y(t)$,经过预加重后的信号即为

$$x(t) = y(t) - 0.95y(t - \Delta t), \tag{5}$$

式中,t 为时间,Δt 为采样间隔.

3)耳蜗模型:采用了由 Patterson 提出的一组 Gammatone 滤波器组来模拟耳蜗的特性[13].在此 Gammatone 滤波器组中,每个通道的 Gammatone 滤波器由 4 个半正交的二阶滤波器级联构成.图 3 为 100—16000 Hz 频率范围内由 25 个 Gammatone 滤波器所构成的耳蜗滤波器组的滤波器响应图.由图可以看出,滤波器在对数轴上的峰值点的分布基本为等间隔分布,这与耳蜗模型的特点相符.根据 Hu 等人对 CASA 系统的研究结果[11],我们对 50—16000 Hz 的频带范围内划分了 128 个通道的 Gammatone 滤波器组,这样能够较地反映此频带内语音的基频和谐波特征.

Gammatone 滤波器组中每个频带的滤波器冲击响应为

$$g(f_c, t) = b^a t^{a-1} \mathrm{e}^{-2\pi bt} \cos(2\pi f_c t), \tag{6}$$

式中,f_c 为中心频率,t 为时间,$a = 4$ 为滤波器阶数.

b 为滤波器衰减因子,它决定了脉冲的衰减速度,与滤波器的带宽有关.耳蜗基底膜对声音信号的不同频率具有非线性选择性,所以滤波器的带宽随着中心频率的升高而增大,可以根据人耳临界频带的等效矩形带宽(ERB)确定,计算公式为

$$\mathrm{ERB}(f) = 24.7 \times \log_{10}\left(\frac{4.37 \times f}{1000} + 1\right). \tag{7}$$

令 $b = 1.019 \times \mathrm{ERB}(f)$,设 $x(t)$ 为输入信号,对于每一个时刻每一个滤波器通道 c,设 f_c 为中心频率.则相应的时频单元 $x(c,t)$ 为

$$x(c,t) = x(t) * g(f_c, t), \tag{8}$$

式中,$x(t)$ 为输入信号,$g(f_c, t)$ 为相应的 Gammatone 滤波器;"$*$"代表卷积,每个通道的输出向后延时 $(a-1)/(2\pi b)$,可以补偿滤波器的延时.

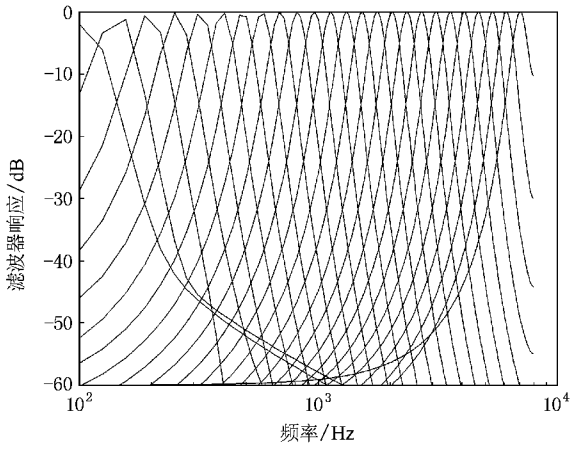

图 3 25 通道 Gammatone 滤波器组响应图

2.2.2 谐波分析

经过模拟耳蜗型的 Gammatone 滤波器处理后的 $x(c,t)$ 即为在时间 t 和通道 f 的时频能量分布,这种分布是符合人耳特性的.下面我们再利用这种时频分布计算符合耳蜗模型的谐波成分.

为了防止不同的能量相关系数标准不同,先将系数进行归一化处理,归一化后的自相关系数为

$$r_{\mathrm{A}}(c,t,\tau) = \frac{x(c,t)x(c,t-\tau)}{x^2(c,t)}, \tag{9}$$

式中,c 为相应的滤波器通道,t 为时刻,τ 为时延.

$$C_{\mathrm{H}}(c,t) = \sum_{\tau=0}^{L} \tilde{r}_{\mathrm{A}}(c,t,\tau)\tilde{r}_{\mathrm{A}}(c+1,t,\tau), \tag{10}$$

式中，L 为所计算的最大时延，人发声的基频一般在 50 Hz 以上，因此 $L = 1000\ \mathrm{ms}/50\ \mathrm{Hz} = 20\ \mathrm{ms}$. $\tilde{r}_{\mathrm{A}}(c,t,\tau)$ 为 $r_{\mathrm{A}}(c,t,\tau)$ 归一化后的结果，$C_{\mathrm{H}}(c,t)$ 即为所得出的考虑到时间和频带上连续性的相关系数.

由于谐波成分之间的时域和频带相关性会大于谐波与噪声的相关性或噪声与噪声的相关性，所以我们通过 $C_{\mathrm{H}}(c,t)$ 的值范围就可以判断在 t 时刻 c 通道的成分是否为谐波成分. 为此我们必须先设定一个阈值，然后比较 $C_{\mathrm{H}}(c,t)$ 与阈值的大小关系来判断谐波成分.

图 4 所示为正常的长元音样本相关系数图，可以看出在一定的通道上显示出强的相关性极值. 而图 5 所示的病变的长元音样本相关系数图，图中的系数显得杂乱而峰值很不明显.

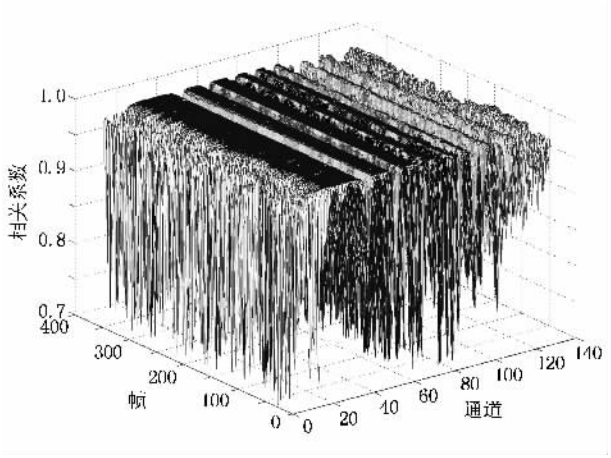

图 4　典型正常长元音噪音样本的交叉相关系数

2.2.3. 谐噪比计算

在对设定一个阈值后，就可以根据这个阈值分离出谐波和噪声成分. 设 $E(c,t) = x(c,t)^2$ 为相应时频块的能量，E_{H} 和 E_{N} 分别为谐波和噪声能量，则

$$\begin{aligned} E_{\mathrm{H}} &= \sum E(c,t), C_{\mathrm{H}}(c,t) > U_{\mathrm{H}}, \\ E_{\mathrm{N}} &= \sum E(c,t), \text{其他}, \end{aligned} \tag{11}$$

式中 U_{H} 为分隔谐波与噪声的阈值. 最后得出的信噪比为

$$\mathrm{HNR} = 10 \times \log_{10}\left(\frac{E_{\mathrm{H}}}{E_{\mathrm{N}}}\right). \tag{12}$$

2.2.4　阈值的设定

为了选定阈值，我们选取了经过评级的典型噪音样本 40 例，其中 G0—G3 各 10 例，每例发的长元音数目为 3 个，共 120 个长元音样本. 手工取其中较为平稳的部分，每部分的长度都在 3 s 以上.

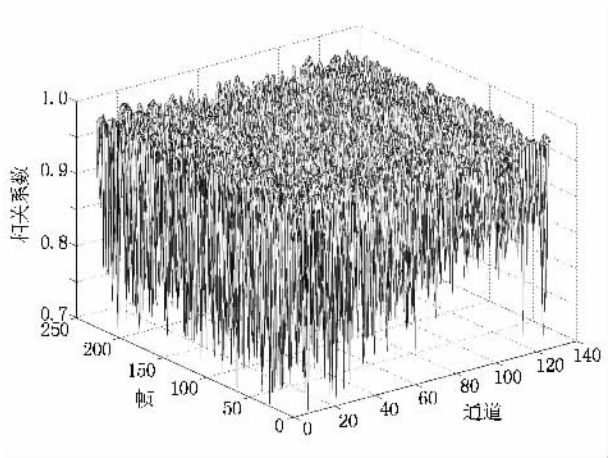

图 5　典型病变长元音样本的交叉相关系数

阈值初始设置的区间是 0.1—0.9，增加步长为 0.1. 由于在 0.1—0.4 的区间内有样本的所有相关系数都在阈值之上，无法得到有效的谐噪比数值. 在 0.5—0.9 区间内的谐噪比与噪音级别的相关度数值（具体见第 3 节）如表 1 所示. 其为单调递减关系.

表 2 所设置的阈值区间为 0.90—0.99，增加步长为 0.01. 相关度为单调递减关系.

表 3 所设置的阈值区间为 0.990—0.999，增加步长为 0.001. 阈值在 0.990—0.993 的范围内，相关度达到了最大值，因此最后选定的阈值大小为 0.9915.

表 1　阈值在 0.5—0.9 的相关度

阈值	0.50	0.60	0.70	0.80	0.90
相关度	−0.108	−0.147	−0.172	−0.620	−0.736

表 2　阈值在 0.90—0.99 的相关度

阈值	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
相关度	−0.736	−0.744	−0.751	−0.758	−0.765	−0.770	−0.776	−0.782	−0.787	−0.798

表 3　阈值在 0.990—0.999 的相关度

阈值	0.990	0.991	0.992	0.993	0.994	0.995	0.996	0.997	0.998	0.999
相关度	−0.798	−0.798	−0.798	−0.798	−0.789	−0.787	−0.784	−0.776	−0.757	−0.707

2.3. 提高元音稳定性

为了提高评估方法的客观性和稳定性,降低偶然性,一些研究者对样本的选择提出了一些准则[14],例如:1)人发声的起始阶段和结束阶段比较不稳定,因此计算结果时应该选择其中较为稳定的部分. 2)参数的数值一般都受到分析窗口位置与长度的影响.因此许多专家建议应该重复测试数据,比如连续进行三次测量,由测试者选则一次测量值或取三次其平均值.但是这种选择方法上存在一些问题:1)进行的人工干预使得结果不够客观;2)每次选择的标准无法统一,结果很难复现;3)费时费力.

为此,下面提出了一种自动进行切分的方法,能够根据能量大小与能量变化率自动切分出长元音中稳定的发音部分,同时根据长度对几组数据的得分进行加权平均,这样能够更客观更有效的对长元音进行分析.

2.3.1. 切分

由于录音环境是在消声室,背景噪音很小,因此切分方法采用了基于能量的活动语音检测(VAD)算法对样本进行第一次切分.在对起始能量和静音长度设定相应的参数后,就能有效地将每次发音都切分开.

2.3.2. 选择稳定部分

先对信号进行去直流成分,然后进行能量归一化.这样可以避免不同大小能量的样本之间切分标准的差异.

取 20 ms 为一帧,帧与帧之间有 50%的叠接,计算出每一帧的能量.将能量变化曲线用鲁棒多项式局部回归方法进行平滑,这种方法通过限定奇异点的权重能有效去除噪点,得出比较平滑的能量曲线.对平滑的能量曲线求差分,得出能量变化率.

如图 6 所示,利用能量曲线和能量变化率曲线和均值(图中横虚线)的距离,去除掉头部和尾部能量变化率较大的部分,得出一段较为平稳的长元音(图中竖虚线间部分).

2.3.3. 多个样本的加权平均

由于人在发持续时间较长的元音时往往稳定性较好,因此有必要对发音时间长的样本采取更大的权重.因此按照切分出的每段的长度比率计算出加权系数,根据加权系数和分段数可以算出最终的 HNR,

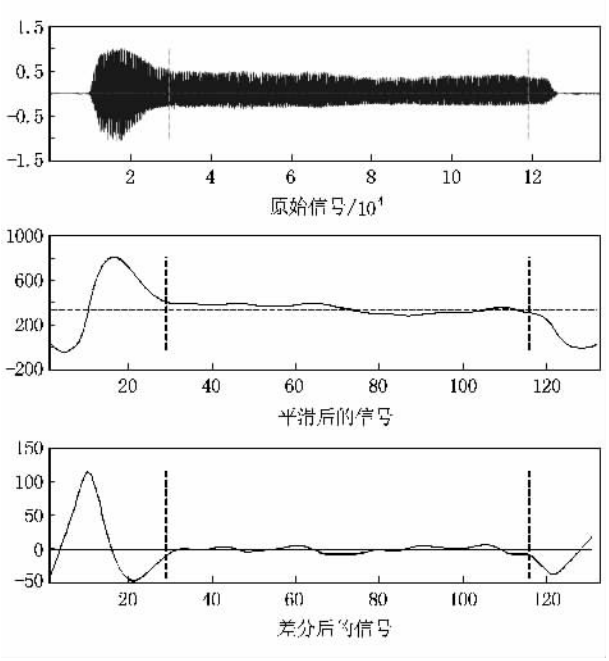

图 6 信号稳定部分的选择

$$\mathrm{HNR}' = \frac{L_n \times \mathrm{HNR}_n}{\sum_{n=1}^{M} L_n}, \tag{13}$$

式中,M 为切分出的样本总数,L_n 为第 n 段样本稳定部分的长度,HNR′ 为最终计算出的谐噪比.

3. 实 验

3.1. 样本选择

实验过程选取了在中国人民解放军总医院(301 医院)就诊的嗓音障碍患者和无嗓音障碍的成年人共 227 名,其中男性 106 名,女性 121 名,年龄分布从 9 岁到 71 岁.每个人发长元音/a/3—5 遍,得出的长元音样本总数为 817 个.

3.2. 样本评级

对病变嗓音程度的评级,试验中采用了 Hirano 所提出的 GRBAS[15] 评级方法中的 G 级,即由嗓音医学专家根据异常嗓音样本的综合嘶哑程度,对样本进行 0—3 级别的打分,其中 0 级为正常,嘶哑程度最小,3 级的嘶哑程度最大.GRBAS 中 G 的评估结果相对更加稳定可靠,现在已逐渐成为病变嗓音评级中最主要的评级方法[4,5,8].

3.3. 相关性分析

与传统的谐噪比计算方法相比较，对于样本集中的 62 例样本，传统方法无法识别出它的基频周期，因此无法得出有意义的结果．而本文提出的基于听觉模型的方法则有效提出了谐噪比数值．

对于得出的谐噪比数值，我们采用了 SPSS 14.0 中的二元相关性分析方法[16] 对数据进行相关度分析(表 4)．

表 4　相关度数值比较

	HNR 与 G 的相关性
传统基频位置方法	－0.620
听觉模型相关方法	－0.790

3.4. 误差分析

表 5 列出了基于听觉模型的计算方法与传统计算方法在不同分级中的平均值、标准差和方差．从表中可以看出，基于听觉模型的计算方法所得出的标准差和方差明显要小于传统的计算方法，因此区分度更好．

表 5　误差分析

	基于听觉模型的计算方法					传统计算方法			
	0 级	1 级	2 级	3 级		0 级	1 级	2 级	3 级
平均值	15.63	10.33	6.75	2.79	平均值	22.45	20.32	17.40	12.15
标准差	2.61	3.23	3.25	4.57	标准差	3.58	3.65	4.19	5.23
方差	6.80	10.43	10.59	20.92	方差	12.83	13.36	17.54	27.38

4. 结　论

本文提出了一种更接近人耳听觉模式的长元音谐噪比计算方法，相对于传统的谐噪比计算方法而言，这种方法更符合人耳的听觉模型，可以更准确的对长元音噪音样本进行谐噪比分析，对于一些无法用传统方法提取基频周期的样本也能有效地进行分析．同时利用样本的能量及能量变化率自动切分出长元音样本中的稳定部分，从而提高了样本的稳定性和区分度．

[1] Yumoto E, Sasaki Y, Okamura H 1984 *J. Speech Hear Res.* **27** 2

[2] Boersma P 1993 *IFA Proceedings* **17** 97

[3] Krom G d 1994 *J. Speech Hear Res.* **37** 985

[4] Yu P, Ouaknine M, Revis J, Giovanni A 2001 *J. Voice.* **4** 529

[5] Bielamowicz S, Kreiman J, Gerratt B R, Dauer M S, Berke G S 1996 *J. Speech Hear Res.* **39** 126

[6] Krom G d 1993 *J. Speech Hear Res.* **36** 254

[7] Yegnanarayana B, d'Alessandro C, Darsinos V 1998 *Speech and Audio Processing, IEEE Transactions on* **6** 1

[8] Zhao S G, Bu F L, H.S Y, Han C C 2003 *International Conference on Natural Language Processing and Knowledge Engineering*

[9] Cheolwoo J, Tao L, Jianglin W 2005 *Engineering in Medicine and Biology Society* 2005 *IEEE-EMBS* 2005. 27*th Annual International Conference of the* 4678－4681

[10] Bergman A 1990 *Auditory scene analysis: the perceptual organization of sound* (Cambridge, MA: The MIT Press.)

[11] Hu G N, Wang D L 2006 *Acoustic Echo and Noise Control* (Berlin Heidelberg: Springer) 485－515

[12] Meddis R, O'Mard L 1008 (Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc.) 43－58

[13] Patterson R D, Nimmo-Smith I, Holdsworth J, Rice P 1988 *MRC Applied Psych.* 2341

[14] Karnell M P 1991 *J. Speech Hear Res.* **34** 544

[15] Hirano M 1981 *Clinical Examination of voice* (Wien: Springer Verlag)

[16] Kirkpatrick L A 2005 *A Simple Guide to Advanced Statistics for SPSS*, Version 13.0 (California: Wadsworth Publishing)

Automatic pathological voice evaluation based on auditory model*

Wang Di[1)2)†] Fu Qiang[2)] Yang Lin[2)] Yu Ping[3)] Yan Yong-Hong[2)] Feng Ji[1)]

1)(*Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*)

2)(*Think IT Speech Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100080, *China*)

3) *Chinese PLA General Hospital, Beijing* 100853, *China*)

(Received 15 July 2007; revised manuscript received 4 December 2007)

Abstract

Sustained vowel analysis is an important method in pathological voice evaluation system. In view of the fact that conventional methods generally analyzes the hand-cut samples in FFT frequency domain, an auditory model is proposed as an accurate and robust method for use in pathological vowel analysis. An automatic sample cutting method increases the stability of analysis. Expertiment showed this method is more effective and robust.

Keywords: voice evaluation, auditory model, sustained vowel analysis, HNR

PACC: 4300, 4370, 8736, 8734

* Project supported by the National Basic Research Program of China (Grant No. 2004CB318106), the National Natural Science Foundation of China (Grant Nos. 10574140, 60535030), the National High Technology Research and Development Program of China (Grant Nos. 2006AA010102, 2006AA01Z195).

† Corresponding author, E-mail: dwang@hccl.ioa.ac.cn