基于听觉模型的耳语音的声韵切分*

丁　慧　　栗学丽　　徐柏龄

(南京大学声学所　近代声学国家重点实验室　南京 210093)

2002 年 11 月 25 日收到

摘要　本文分析了耳语音的特点，并根据生理声学及心理声学的基本理论与实验资料，提出了一种利用听觉模型来进行耳语音声韵切分的方法。这种适用于耳语音声韵切分的听觉感知模型主要分为四个层次：耳蜗对声音频率的分解机理；听觉系统的时域和频域非线性变化；中枢神经系统的侧抑制机理。这种模型能反映在噪声环境下人对低能量语音的听觉感知特性，因而适于耳语音识别，在耳语音声韵母切分实验中得到了满意的结果。

关键词　耳语音，听觉模型，声韵切分

Initial/final segmentation of Chinese whispered speech based on the auditory model

DING Hui　　LI Xue-Li　　XU Bo-Ling

(State Key Laboratory of Modern Acoustic, Institute of Acoustics, Nanjing University, Nanjing 210093)

Abstract　In this paper, the characteristics of whispered speech are discussed, and a new approach for initial/final segmentation of Chinese whispered speech is proposed on the basis of psychological acoustic theories and experiments. With the mainly four levels of signal processing, this model can represent human's perceptual features of low energy speech, so it is more suitable for the whispered speech recognition. With the experiments of the division between the initial and the final of whispered speech included 386 Chinese syllables at 5dB SNR, the results show that the proposed approach can catch the features of whispered speech more accurately.

Key words　Whispered speech, Auditory model, Initial and final segmentation

1　引言

耳语是一种常见的但不同于正常语音的低声级的发音模式。在耳语发音时，声带几乎没有振动，音量小，但这不妨碍人们对耳语音的理解及交流。随着移动手机的广泛使用，人们常常需要在公共场合进行通话，为了不影响其他人或者保证通话的保密性而使用耳语音，这就不仅要求手机有很好的接收灵敏度而且要具备识别耳语声 [1] 和将耳语声转化为正常语音后再进行传输的能力 [2] 。耳语音的声韵分割在耳语音识别及转化中有重要的作用。如为了提高耳语音的识别率，须进行声韵分割。在耳语转化为正常音时，需确定哪一部分是韵母，

* 国家自然科学基金资助项目 (60272037)

以便转化时在韵母部分加上基音。

正常语音的声韵切分方法主要有：过零率、短时能量、线性预测编码 LPC 参数法、倒谱参数法以及基于小波变换的切分方法等。由于耳语音的韵母部分不像正常韵母那样是准周期的，所以以往的基于准周期性的正常语音清浊音切分法已不适用于耳语音的声韵切分。

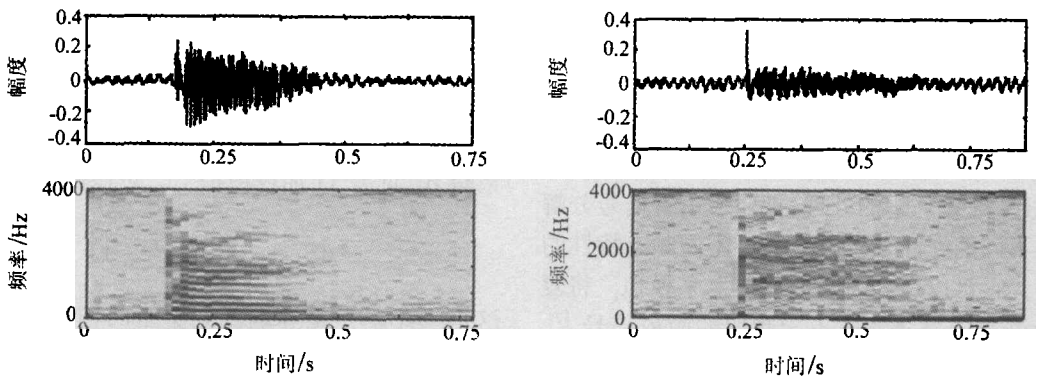
目前国内外对耳语音这方面的研究还很少，文献 [2] 中提到了一种耳语音的浊音的恢复方法：在混合激励的线性预测模型 (MELP) 中，固定的认为四个低频率带的语音为浊音段，高频带的语音段为清音段。由于没有进行声韵分割，声母部分也添加了基频，恢复出的正常音产生失真。至今还没有找到涉及耳语音声韵分割算法的相关文献。大部分研究者对于耳语音的研究集中在：耳语音高的感知、耳语音与正常语音的比较。本文提出了一种基于听觉模型的耳语音声韵切分方法，若干研究表明，在语音识别系统中考虑到听觉系统的处理特点而引入一些预处理，会提高整个系统的识别率。听觉系统还有抑制噪声的作用，这可满足耳语音在低信噪比下的处理要求。实验结果证实了用此法可实现耳语音的声韵分割。

2 耳语声的声学特性

一般说来，由声带振动产生的音统称为浊音 (Voiced)，而声带不振动产生的音统称为清音 (Unvoiced)。在耳语发生过程中，声门半开，呼出的气体通过声门开口的收缩产生湍气流，当气流速度与声道的横截面积之比大于某个门限时，便产生完全由噪声激励的耳语音。声带是不振动的，也就是说耳语音是没有基频的，耳语的韵母对比于正常语音来说，不是准周期的。但是耳语音的共振峰仍然存在。其变化表现为，耳语音的第一共振峰的幅值较低，位置向高频偏移，带宽变宽。有研究表明，虽然耳语音没有基频，但是由耳语的振幅特性可感知的音高仍然存在 [3]。在同样的发音内容情况下，耳语音的能量平均比正常语音低 20dB 左右。从频谱上可见正常语音的韵母与耳语音的韵母差别很大，而声母却没有明显的差异。对识别感知起重要作用的耳语信号的频率段约在 500–4000Hz 之间。图 1 是正常语音与耳语音 /pa/ 的波形图和语谱图的比较。

3 基于听觉感知模型的耳语音声韵切分法

人耳的听觉系统具有计算机识别系统望尘莫及的抗噪性能和智能处理能力，这吸引了众多的研究者从事人耳听觉模型的研究。到目前已经有了几种较具代表性的人耳听觉模型和语音信号的听觉表示方法：



(a) /pa/ 正常语音　　(b) /pa/ 耳语音

图 1　正常语音与耳语音的波形图和语谱图比较

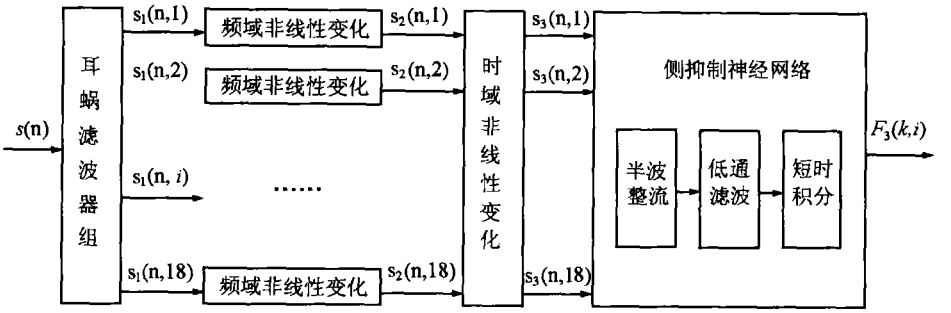


图 2　耳语听觉感知模型

(1) 时域听觉模型 [4], 这种模型研究了声音信号序列在听觉系统里随时间变化产生的频率、相位、幅度等变化; (2) 频域听觉模型 [5], 该模型分为三个部分: 基底膜滤波阶段、内毛细胞的非线性表达、自动增益控制阶段。总的来说, 这类听觉模型重点在于表达整个听觉系统在最佳频率或特征频率处的神经发放; (3) 时域和频域相结合的听觉模型 [6], 这种模型强调信号在听觉系统中的时空分布及反应, 和听神经对声音信息的处理与传输。

在耳语信号中, 存在着时频偏移, 在满足一定的信噪比的情况下, 人耳可以轻松的进行理解辨别, 包括声母与韵母的切分。本文在以上三种模型的基础上, 根据心理声学和生理声学研究听觉感知特征, 并结合人耳对耳语音感知特征的实验数据建立了相应的耳语听觉感知模型。以下分成四个层次进行耳语听觉感知模型的建模并进行耳语音的声韵切分。图 2 是整个耳语听觉模型的框图。

3.1　耳蜗滤波 — 基底膜的频率分解

对人耳的听觉系统研究表明, 声波传入到耳蜗后, 在基底膜引起振动并以行波的方式向前传递, 通过耳蜗的处理, 把时域语音信号分解成在不同的空间轴位置上具有不同频率特性的信号。它可以用一组带通滤波器来模拟, 将复杂的语音信号分解成不同频段的简单信号 [7], 便于进一步的分析和处理。在 20~16000Hz 的听阈范围内的频率可分成 24 个频率群, 这种临界频带用 Bark 标度表示, 与频率 f 的关系见表达式 (1)。本文中, 语音信号经过 8KHz 采样并经过 4KHz 的低通滤波器滤除去高频分量。模型中的耳蜗滤波器组以 Bark 标度划分。利用 Yule-Walker 函数设计了 18 个 IIR 带通滤波器, 在模型中, 耳蜗滤波器组输出表示为 $s_1(n,i)$。滤波器组的频响特性以 Zwicker 给出的 "激励强度-临界频带率模型" 为依据 [8]。

$$z(\text{Bark}) = 13 \cdot \arctan[0.76 f(kHz)] + 3.5 \cdot \arctan[f(kHz)/7.5]^2 \qquad (1)$$

3.2　声音信号的频率非线性变化 — 听觉灵敏度加权

人耳对于声音强还是弱的判断, 是与声音的强度有关的, 也与声音的频率有关, 用声音的响度级 P, 单位为方 (Phon), 来表示人耳对于频率不同的纯音的听辨灵敏度。响度是一种主观心理量, 是人类主观感觉到的声音强弱程度。一般来说, 当声音频率一定时, 声强越强则响度也就越大。相同的声强, 频率不同时响度也可能不同。由于听觉系统这种特性对不同的频率的刺激的灵敏度不同, 模型中以灵敏度系数 α_i(见表 1) 加权模拟 [7], 见 (2) 式。

$$s_2(n,i) = \alpha_i * s_1(n,i) \qquad (2)$$

这里 $s_1(n,i)$ 是耳蜗滤波器组的输出, $i = 1,\cdots,18$, $n = 1,2,\cdots L$, L 为耳语音序列的长度。$s_2(n,i)$ 为频率非线性变化级的输出。图 3 是听觉模型中模拟的听阈灵敏度加权后的耳蜗滤波器组的特性曲线。横坐标表示归一化频率

表 1　灵敏度系数 α_i 加权

临界频带 (Bark)	1	2	3	4	5	6	7	8	9
α_i (dB)	27.00	15.67	10.67	7.33	4.67	3.11	2.22	2.00	2.00
临界频带 (Bark)	10	11	12	13	14	15	16	17	18
α_i (dB)	2.00	1.89	1.25	0.00	−1.60	−3.20	−4.65	−4.32	−1.60

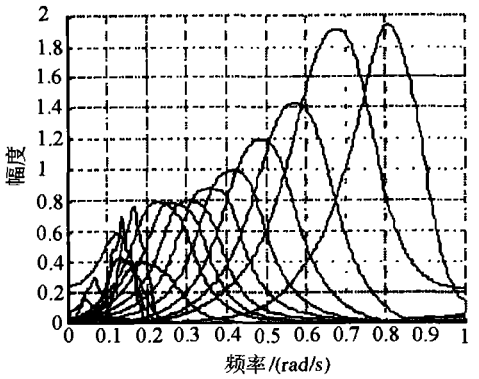


图 3　听阈灵敏度加权后的耳蜗滤波

3.3　听觉系统内的时域非线性变化

根据听觉系统的非线性饱和特性：随声音刺激强度增加，发放率随之增加，当声音强度达到一定水平时，激发率达到饱和。由于耳语声是弱刺激，可以认为发放率随声音刺激强度线性变化。这里的时域非线性变化级是为了模仿听觉神经纤维的锁相程度和激励源的密度之间的关系 [10]，在听觉系统中常用一个非线性单调递增的函数来表示：

$$g(x) = \lg(1 + x) \tag{3}$$

x 在模型中 (图 2) 为 $s_2(n,i)$。这个非线性函数实现了振幅的非线性压缩，可以使系统对噪声非常敏感，具有较好的鲁棒性 [11]。听觉中的掩蔽效应是耳语听觉非线性的重要功能，当两个响度不等的声音作用于人耳时，响度较高的频率成分的存在会影响到对后来响度较低的频率成分的感受，使其变得不易觉察。对于耳语声来说，噪声的影响很大，而噪音的存在同样也会对耳语音产生掩蔽。耳语音是强度非常小的语音信号，在高阈值和低增益的状态下，耳语声这种弱的刺激信号常常低于感知阈值，而不被感知，人耳常常可以提高注意力，来调节对弱音的感知。我们采用文献 [7] 连续语音的非线性滤波来模拟前置掩蔽和后向掩蔽特征，这一级输出用 $s_3(n,i)$ 来表示。图 4 例示了脉宽分别为 1ms、3ms、10ms、30ms、100ms、200ms 的脉冲信号作为激励信号时，由于掩蔽效应而引起的在时间上前后沿变化的情况。

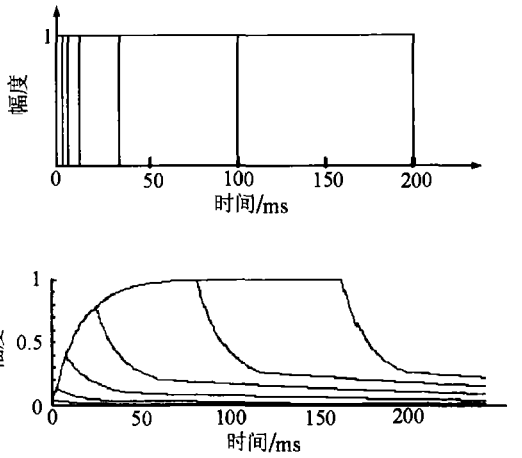


图 4　不同脉冲信号及前置掩蔽和后向掩蔽的作用

3.4　侧抑制神经网络[12]

在一般情况下耳语音的信噪比相对于正常语音较低，为了提高识别能力，听觉模型中引入侧抑制神经网络。许多研究表明，侧抑制神经网络能够增强输入模式的对比，突出波峰和边缘，这对耳语识别而言是很有益的 [6]。其中分成三个部分。

(1) 半波整流

沿基底膜分布的毛细胞处于基底膜和耳蜗覆膜之间，基底膜的振动造成基底膜与耳蜗覆

膜之间的剪切运动，引起毛细胞的去极化和超极化，使得内毛细胞对流体的振动速度相应产生半波整流效应。激发率几乎是刺激波形正的部分的复本，而负的部分不发放。为反映内毛细胞的半波整流特性，模型中只取正的响应值。

$$F_1(n,i) = \max[s_3(n,i), 0] \tag{4}$$

(2) 低通滤波

毛细胞会对声音信号产生时间平滑作用，这里用一个二阶 IIR 型数字滤波器 $h_{lp}(n)$ 加以实现，其时间常数取 1.3ms[7] 。

$$F_2(n,i) = F_1(n,i) * h_{lp}(n) \tag{5}$$

(3) 短时积分

听觉神经中枢不像神经纤维那样能及时跟上声音刺激，一般反应时间为 10~20ms，系统的输入信号分成 10ms 一个语音帧，即每帧 80 个点，用短时 (一个语音帧) 积分来模拟这一过程。那么侧抑制神经网络最后的输出为

$$F_3(k,i) = \int_n^{n+T} F_2(n,i)dn \tag{6}$$

$T = 80$ (10ms), $i = 1, 2, \cdots 18$, 表示 18 个临界滤波器滤波序列， $k = 1, 2, \cdots L/T$, 表示语音帧序列。

3.5 基于听觉感知模型的耳语音声韵切分

正常语音的声韵切分方法主要有：过零率、短时能量、线性预测编码 LPC 参数法、倒谱参数法以及基于小波变换的切分方法等。由于耳语音的韵母部分不像正常韵母那样是准周期的，所以以往的基于准周期性的正常语音清浊音切分法已不适用于耳语音的声韵切分。如图 5(b) 韵母的能量与声母的能量没有明显的差别， (c) 声母的过零率与韵母的过零率几乎相同。

声母音长比较稳定，不太因人而异，只要识别出声母，那么就可以很容易地进行声韵切分了。下面是整个声韵切分的信号处理过程：

(1) 输入信号 $s(n)$ 经过非线性滤波，分成 18 个 Bark 的信号 $s_1(n,i)$, $i = 1, \cdots, 18$; $n = 1, 2, \cdots L$, L 为耳语音序列的长度；

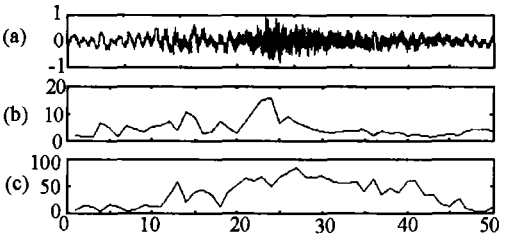


图 5 耳语音 /sai/ 的时域图 (SNR=5dB)。横轴表示语音帧数，纵轴表示： (a) 幅度 (b) 短时能量 (c) 过零率

(2) 听觉灵敏度加权

$$s_2(n,i) = \alpha_i * s_1(n,i) \tag{2}$$

(3) 时域非线性变化见公式 (3)，连续语音的非线性滤波来模拟前置掩蔽和后向掩蔽特征，此级输出为 $s_3(n,i)$

$$g(x) = \lg(1+x) \tag{3}$$

(4) 半波整流

$$F_1(n,i) = \max[s_3(n,i), 0] \tag{4}$$

(5) 低通滤波

$$F_2(n,i) = F_1(n,i) * h_{lp}(n) \tag{5}$$

(6) 短时积分， $T = 80$ (10ms), $i = 1, 2, \cdots 18$, $k = 1, 2, \cdots L/T$,

$$F_3(k,i) = \int_n^{n+T} F_2(n,i)dn \tag{6}$$

(7) 切分参数

$$F(k) = \mathrm{mean}[F_3(k,i)] \tag{7}$$

这里取每帧中听觉模型的 18 个临界频带输出值的平均值 $F(k)$ 来作为声韵切分的参数， k 表示语音帧序列。由于语音信号经过归一化处理，所以可以统计得到统一的阈值 F_0(误差范围 1-3 帧), 实现声韵切分。

表 2　耳语音听觉模型声韵切分实验统计

声母	不送气塞音			送气塞音			鼻音		边音	卷舌音	零声母		
	b	d	g	p	t	k	n	m	l	r	y	w	无
测试样本数	15	21	20	17	20	18	19	19	23	14	12	9	6
正确切分数	14	19	18	15	17	15	17	17	21	13	9	6	4

声母	清擦音					不送气塞擦音			送气塞擦音		
	s	x	f	sh	h	z	zh	j	ch	c	q
测试样本数	16	14	9	18	19	17	19	14	15	18	14
正确切分数	14	11	8	16	16	14	17	11	11	17	12

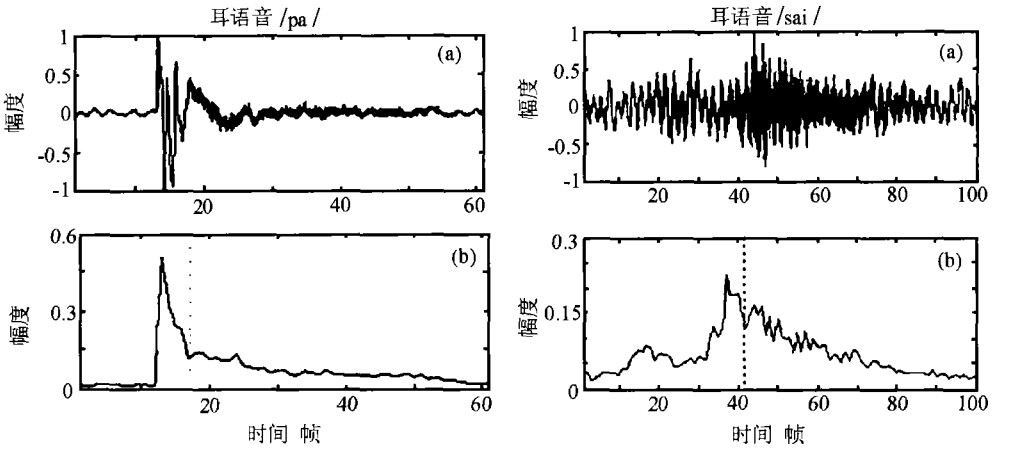


图 6　声韵切分实验举例　(a) 耳语音波形　(b) 切分参数曲线 (虚线表示切分线)

4　实验

为了检测该声韵切分算法的性能，我们对汉语普通话耳语音 386 个音节进行了声韵切分实验，其中 23 个声母，33 个韵母。实验环境是普通实验室，有空调、计算机、日光灯、还有人走动等背景噪声，平均信噪比 5dB 左右。采样率 8kHz, 帧时长取 10ms 。表 2 给出了声韵切分结果的统计数据。

将计算切分结果与人工切分结果相比较，误差在 1~3 帧之内 (每帧 80 个数据点)，得到平均切分正确率为 85.6%，切分的错误主要表现在两个方面：(1) 零声母的类似韵母性质，听觉模型不能很好的反映零声母与韵母之间的能量突变；(2) 介音为 /i/ 和 /u/ 音节的声韵母界限不明显，变化规律多变，导致算法在辨别含有这类音节时错误率较大。图 6 给出了以塞音开头的耳语音 /pa/ 和以清擦音开头的 /sai/ 的声韵切分实验波形结果。

5　结论

本文采用的适合于耳语音声韵切分的方法反映了耳语音的过渡特征，过渡特征除了指声母与韵母之间的过渡外，还包括前后音节间的相互影响，其中的非线性变化级，特别是掩蔽效应能较好的描述这种过渡特征；在一般环境中，耳语音的信噪比往往很小，噪声影响很大，本方法的时间非线性变化级和侧抑制神经网络级都具有很好的噪声抑制能力，这样就大大增加了识别率，因为它反映了在噪声环境下人对低能量语音的听觉感知特征。我们把它用于耳语音的声韵母切分，取得了较好的结果。

(下转第 44 页)

目标进行左右舷分辨。该方法对单频和信号相对带宽不太大的宽带情况均有效，而如果频率与预先设定频率偏差太大或信号相对带宽太大时，则左右舷分辨性能急剧下降，无法使用。

若要将该方法实际应用，必须保证 $l \leq 0.5\lambda$，最佳的间距是 $l = 0.25\lambda$，这样既能有较大的左右舷分辨增益，指向性函数又不会出现多值区间。为提高处理增益，可在波束形成前，将接收信号进行带通滤波。

此外，本文的部分结论也可扩展到三元水听器组：基于几何相移模型的三元水听器组也可在带宽不太大时完成对目标的左右舷分辨。

参 考 文 献

1 Jean Bertheas, Gilles Moresco, Philippe Dufourco. Linear hydrophonic antenna and electronic device to remove right/left ambiguity, associated with the antenna. United States Patent, Patent Number: 5058082, Oct.15, 1991.

2 Doisy Y. Port-starboard discrimination performances on actived towed array systems. UDT95, France, 125～129.

3 杜选民，朱代柱，赵荣荣等. 声学学报， 2000, **25**(5): 395～402.

(上接第 14 页)

振”是复合振动系统中存在的一种普遍现象。

参 考 文 献

1 范国良，应崇福，林仲茂等. 应用声学， 1982, **1**(1):2～7.

2 应崇福，范国良. 应用声学， 2002, **21**(1):19～25.

3 马玉英，丁大成. 物理学报， 1987, **36**(2):208～215.

4 林书玉，张福成. 声学学报， 1992, **17**(6):251～255.

5 周光平，鲍善惠，程存第. 应用声学， 1994, **13**(6):39～42.

6 鲍善惠. 应用声学， 1998, **17**(4):6～10.

7 丁大成，马玉英. 陕西师大学报， 1985, **2**(2):23～25.

8 Fan Guoliang, Zhang Weixian. Proc International Conference on Ultrasonic Technology, Toyohashi, 1987. 323～328.

9 赵继，王立江，孟继安. 声学学报， 1992, **17**(1):22～29.

10 赵波，何定东. 机械工艺师， 1998, (6):4～6.

11 林仲茂著. 超声变幅杆的原理和设计. 北京：科学出版社, 1987.

(上接第 25 页)

参 考 文 献

1 Taisuke Itoh, Kazuya, Fumitada. *ICASSP*, 2002, 389–392.

2 Morris R W, Clements M A. The 2nd Int. Workshop MAVEBA, 2001.

3 Higashkawa M, Nakai K *et al*. *J. of Voice*, 1996, **10**:155～158.

4 David J M Robinson, Malconlm J Hawksford. The 107th Conference of the Audio Engineering Society, New York, September 1999.

5 Cosi P, Pasquin S *et al*. ICSLP'98 Proceedings, Page (NA), 1053.

6 Yang Xiaowei, Wang Kuansan *et al*. *IEEE Trans. on Information Theory*, March 1992, **38**(2):824～839.

7 赵鹤鸣，周旭东. 电子科学学刊， 1994, **16**(5):513～517.

8 Scharf B. Bands C. in Tobias J V (cd.). New York: Academic Press, 1970. 159～202.

9 杨行峻，迟惠生等. 语音信号数字处理. 北京：电子工业出版社， 1995. 4～33.

10 Doh-suk Kim, Soo-Young Lee, Rhee M Kil. *IEEE Trans. Speech Audio Proc.*, 1999, **7**(1):55～69.

11 李鸣华. 计算机与现代化， 2000, **67**:9～13.

12 戴明扬，余凯等. 应用声学， 2001, **19**(2):121～126.

(上接第 48 页)

参 考 文 献

1 何祚镛著. 结构振动与声辐射，第 1 版. 哈尔滨：哈尔滨工程大学出版社， 2001.

2 郑士杰，袁文俊，缪荣兴等编著. 水声计量测试技术，第 1 版. 哈尔滨：哈尔滨工程大学出版社， 1995.

3 Read B E, Dean G D. 聚合物和复合材料的动态性能测试，第 1 版. 上海：上海科技文献出版社， 1986.

4 缪荣兴，宫继祥编. 水声无源材料技术概要，第 1 版. 杭州：浙江大学出版社， 1995.

5 王荣津编. 水声材料手册，第 1 版. 北京：科学出版社, 1983.