第14卷 第1期
1992 年 1 月

电 子 科 学 学 刊
JOURNAL OF ELECTRONICS

Vol. 14, No. 1
Jan., 1992

听觉模型及其应用*

杨 俊 樊昌信
(西安电子科技大学信息工程系,西安 710071)

摘要 结合生理声学和心理声学资料,本文提出了一个由非均匀间距带通滤波器组、检测器组和主频选取机构等三部分组成的听觉模型.它们依次表征基底膜、内毛细胞和神经纤维的特性. 基于所建听觉模型并结合修正的临界带宽参数构成的语音分析系统, 输入模拟了鼓膜上的声压波,输出模拟了各种神经冲动图特征. 语音综合系统采用简单相加法来获取重建语音. 计算机模拟实验表明,重建语音是高可懂的、自然的, 证明了所建听觉模型的正确性以及临界带宽参数的修正是有意义的.

关键词 听觉系统;临界带宽;语音分析/综合

一、听觉模型的构成、理论基础和参数的确定

1. 听觉模型的构成和理论基础

我们建立的听觉模型结构如图 1 所示.

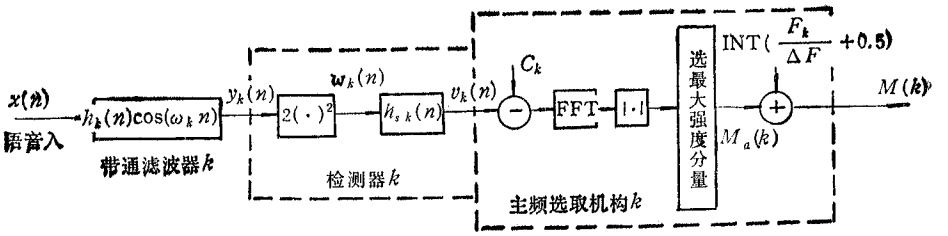


图 1 听觉模型的第 k 个通道 ($k=1,2,\cdots,K$)

图 1 中, C_k 为 $v_k(n)$ 的平均值,即 $C_k=\frac{1}{L_0}\sum_{n=1}^{L_0}v_k(n)$, L_0 为帧长度; F_k 为第 k 个通道的中心频率 (Hz), ω_k 为其数字频率 (rad), 即. $\omega_k=2\pi F_k T$, T为时间抽样周期 (s); ΔF 为频率精度 (Hz), 即每两点的频率间隔为 $\Delta F=1/(NT)$, N为 FFT 长度; $M_a(k)$ 和 $M(k)$ 分别为第 k个通道所选出的位于零频率和中心频率附近的频率点(即代表 $M_a(k)\cdot\Delta F$ 和 $M(k)\cdot\Delta F$(Hz)), 称 $M(k)$为第 k 个通道的主频频率;INT(·)表示取整函数.

(1) 带通滤波器 带通滤波器用来描写耳蜗的机械滤波,即基底膜所起的作用.其

1990.11.17 收到,1991.03.15 定稿.
* 国家自然科学基金资助项目.

脉冲响应为 $h_k(n)\cos(\omega_k n)$，$h_k(n)$ 为窗函数．根据文献[1]中基底膜模型，我们可推导出窗函数的形式为

$$h_k(n)=\begin{cases}\alpha_k^3 n^2 T^3 \exp(-\alpha_k nT), & n\geqslant 1\\ 0, & \text{其它}\end{cases}\quad (k=1,2,\cdots,K) \tag{1}$$

这里，α_k 是一个无量纲的正实常数．每个窗单边带带宽为 $0.509\,\alpha_k$(rad/s)．这样，每个带通滤波器的带宽为 $1.018\,\alpha_k$(rad/s)．

(2) 检测器　对输入信号 $x(n)$ 进行带通滤波可看成是一个调幅过程．为了将已调信号 $y_k(n)$ 转换成一个更有用的形式，我们需要对之进行解调，即用检测器进行解调．在这里，检测器用来获取信号的神经表示 (neural representation)，它由半波平方律无记忆非线性器件（用来描写内毛细胞的半波整流特性[2]）和低通平滑滤波器 $h_{sk}(n)$（用来提取神经响应的包络）构成．

无记忆非线性器件有平方律器件、半波平方律器件和半波逐段线性器件等几种．然而，其中具有半波平方律器件的带通滤波器/检测器（简称 F/D）输出可以较好地[3]模拟听觉神经冲动图，因而我们选用的是半波平方律器件．又由于平方律器件输出的平滑形式(即输出包络)与半波平方律器件输出包络完全相同，两者只差一个乘积因子 2[3]，因此，在我们设计的听觉模型中用平方律器件与一个乘法器的级联来代替半波平方律器件．

当然，无记忆非线性器件也可以采用半波逐段线性器件[4]、指数整流器[4]和双曲正切半波检测非线性器件[6]，但本文基本思想不变．采用上述器件并且确定哪种器件更适于描写内毛细胞的整流特性，也是我们今后的一个研究工作．

低通平滑滤波器 $h_{sk}(n)$ 作为包络成分提取器不一定要求与窗函数 $h_k(n)$ 具有同样的形状，但是，为了便于计算机模拟，即用广义短时傅里叶变换 (GSTFT) 的模平方来实现 F/D，则要求其单边带宽 ω_{sk} 与带通滤波器带宽 $2\,\omega_{bk}$ 相等．又由于具有半波平方律器件的 F/D 输出可以用来模拟听觉神经冲动图，而负的冲动率是没有意义的，因而要求 F/D 输出 $v_k(n)$ 为非负的．由于 $w_k(n)\geqslant 0$，要使得 $v_k(n)\geqslant 0$，则要求平滑滤波器的特性满足 $h_{sk}(n)\geqslant 0$．

(3) 主频选取机构　主频选取机构为一个非线性相对谱强度测量．它观察内毛细胞整流输出的包络，忽略神经冲动的细节情况，选择每个通道位于零频率附近的最大强度频率点 $M_a(k)$，进而获得每个通道的主频频率 $M(k)$，这相当于较高层次听觉系统的处理．

2. 听觉模型参数的确定

由生理声学(研究听觉器官的科学)资料我们建立了听觉系统的模型结构．尽管我们还可由生理声学资料导出模型参数，但保证不了所得模型能够模拟人的主观感觉[3]，因此我们还需用其它知识来确定恰当的听觉模型参数．心理声学（研究声音的主观感觉与客观参数间关系的科学）为我们研究听觉系统提供了帮助．本文将利用掩蔽的临界带宽 (critical bandwidth)[7,8]概念来确定所建听觉模型的参数．

表 1 列出了设计一组临界带宽滤波器所必需的参数[7]；文献[8]则给出了临界带宽与中心频率满足表 1 的关系式．可见，17 个滤波器覆盖了 100Hz～4.4kHz 的频率范围，它们的中心频率是不等间隔的，且双边带宽 BW_k(Hz) 随中心频率 F_k 增加而加宽．因

此(1)式中的 α_k 可由表 1 通过下列带通滤波器带宽（rad/s）式子求得

$$1.018\alpha_k = 2\pi \cdot BW_k$$

即

$$\alpha_k \approx 6.17 \times BW_k,\ k = 1,2,\cdots,K \tag{2}$$

带通滤波器中心频率的数字频率（rad）也由表 1 得

$$\omega_k = 2\pi T \times F_k,\ k = 1,2,\cdots,K \tag{3}$$

另外，文献[9]对临界带宽与中心频率的关系式[8]进行了修正．根据这个修正式，我们设计了一组修正的临界带宽滤波器参数，如表 2 所示．可见，27 个滤波器覆盖了 20Hz ～ 4.4kHz 的频率范围，它们的中心频率也是不等间隔的．参数 α_k 和 ω_k 之值也同样按(2)和(3)式由表 2 求得．

表 1　临界带宽滤波器参数

滤波器数目 k	中心频率 F_k(Hz)	临界带宽 BW_k(Hz)	频率范围 (Hz)
1	150	100	100—200
2	250	100	200—300
3	350	100	300—400
4	450	110	400—510
5	570	120	510—630
6	700	140	630—770
7	840	150	770—920
8	1000	160	920—1080
9	1170	190	1080—1270
10	1370	210	1270—1480
11	1600	240	1480—1720
12	1850	280	1720—2000
13	2150	320	2000—2320
14	2500	380	2320—2700
15	2900	450	2700—3150
16	3400	550	3150—3700
17	4000	700	3700—4400

3. 用 GSTFT 的模平方来实现 F/D

输入信号 $x(n)$ 的 GSTFT 定义为

$$X_n(\exp(j\omega_k)) = \sum_{m=-\infty}^{\infty} x(n-m)h_k(m)\exp[-j\omega_k(n-m)] \tag{4}$$

其中 ω_k 为第 k 个通道的中心频率．由于

表 2 修正的临界带宽滤波器参数

滤波器数目 k	中心频率 F_k(Hz)	临界带宽 BW_k(Hz)	频率范围 (Hz)
1	35	30	20—50
2	65	30	50—80
3	100	40	80—120
4	140	40	120—160
5	180	40	160—200
6	230	50	200—250
7	280	50	250—300
8	340	60	300—360
9	400	60	360—420
10	470	70	420—490
11	550	80	490—570
12	640	90	570—660
13	730	100	660—760
14	840	110	760—870
15	960	120	870—990
16	1090	140	990—1130
17	1240	150	1130—1280
18	1400	170	1280—1450
19	1580	190	1450—1640
20	1790	220	1640—1860
21	2020	240	1860—2100
22	2270	270	2100—2370
23	2570	290	2370—2660
24	2890	350	2660—3010
25	3270	400	3010—3410
26	3700	460	3410—3870
27	4200	530	3870—4400

$$\mathrm{Re}[\exp(j\omega_k n)X_n(\exp(j\omega_k))] = \sum_{m=-\infty}^{\infty} x(n-m)h_k(m)\cos(\omega_k m)$$

为带通滤波器 $h_k(n)\cos(\omega_k n)$ 的输出 $y_k(n)$ (Re(·) 为取实部的函数)，因而可用 GSTFT 来实现 F/D 系统的带通滤波器。

又，检测非线性输出 $2y_k^2(n)$中除了一个低频成分 $|X_n(\exp(j\omega_k))|^2$ 外，其余为高频成分（$2\omega_k$ 分量）. 当低通平滑滤波器 $h_{sk}(n)$ 的单边带宽 ω_{sk} 满足下列关系

$$\left.\begin{aligned} 2\omega_{hk} \leqslant \omega_{sk} < 2\omega_k - 2\omega_{hk} \\ 2\omega_{hk} \leqslant \omega_{sk} < 2\pi - 2\omega_k - 2\omega_{hk} \end{aligned}\right\} \tag{5}$$

时（ω_{hk} 为窗函数 $h_k(n)$ 的单边带宽，且 $\omega_{hk} = \pi T \cdot BW_k$），F/D 系统则输出 $|X_n(\exp(j\omega_k))|^2$，即

$$v_k(n) \approx |X_n(\exp(j\omega_k))|^2,\ k = 1,2,\cdots,K$$

因此，为了便于计算机模拟，我们可以用 GSTFT 的模平方来实现 F/D. 上面的近似式是由于 $h_k(n)$ 和 $h_{sk}(n)$ 是可实现的，但不是理想的滤波器的缘故；(5)式中第二个关系式是由离散时间函数的周期谱特性所确定的.

由上述推导过程可见，用 GSTFT 的模平方来实现 F/D 也存在一个缺点，即检测非线性器件必须是平方律器件，这使设计的灵活性受限. 当然，对于我们的听觉系统模拟工作，选择平方律器件是合理的.

为了简化问题，我们选择 $h_{sk}(n)$ 的单边带宽 ω_{sk} 等于 $h_k(n)$ 单边带宽的两倍，即 $\omega_{sk} = 2\omega_{hk}$，这样，(5)式就成为 $2\omega_{hk} < \omega_k < \pi - 2\omega_{hk}$，用表 1 或表 2 中的 BW_k 和 F_k 则表示成 $BW_k < F_k$ 且 $BW_k + F_k < 1/(2T)$. 可见，采用表 1 情况下，当采样速率 $1/T$ 为 10 kHz 时，通道数目 $K = 17$；$1/T = 8$ kHz 时，$K = 16$；采用表 2 情况下，$1/T = 10$ kHz 时，$1/K = 27$；$1/T = 8$ kHz 时，$K = 25$.

进一步，将(1)式代入(4)式，我们推导出 $X_n(\exp(j_k\omega))$ 的递归形式为

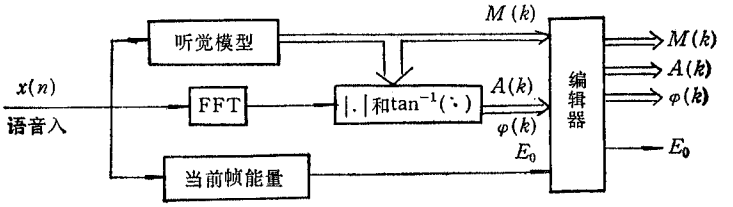
$$X_n(\exp(j\omega_k)) = \sum_{m=1}^{3} P_k(m) X_{n-m}(\exp(j\omega_k)) + \sum_{r=1}^{2} q_k(r) x(n-r) \exp[-j\omega_k(n-r)] \tag{6}$$

其中

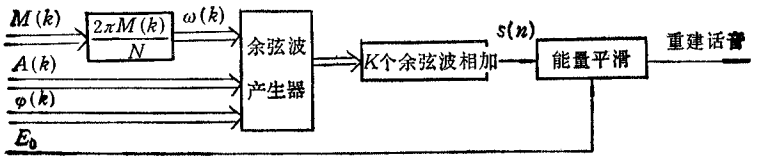
$$q_k(1) = \alpha_k T \exp(-\alpha_k T),\quad q_k(2) = \alpha_k T \exp(-2\alpha_k T),$$

$$P_k(1) = 3\exp(-\alpha_k T),\quad P_k(2) = -3\exp(-2\alpha_k T),\quad P_k(3) = \exp(-3\alpha_k T)$$

(6)式为我们的计算机模拟工作提供了极大方便.



(a) 分析端



(b) 综合端

图 2　语音分析/综合系统结构

二、听觉模型的应用

根据所建听觉模型(图 1),我们设计的语音分析/综合系统如图 2 所示.

分析端，每帧语音首先由听觉模型选出K个主频频率 $M(k)$，进而从输入语音谱中获得相应的K个幅值 $A(k)$ 和K个相位 $\varphi(k)$. 将这K个主频成分和帧能量作为传输参数,故每帧传输参数为 $3K+1$ 个.(利用人耳对相位不敏感的特点,我们还可以不传输相位[10]，这样，每帧传输参数为 $2K+1$ 个.)值得一提的是,本文不涉及编码方案的设计,故分析端最后为"编辑器".

综合端,首先由所接收的 $M(k)$ 按下列公式求出数字频率 $\omega(k)$ (rad):

$$\omega(k)=\Omega(k)\cdot T=2\pi[M(k)\cdot\Delta F]T=2\pi M(k)/N\ (\text{rad}),k=1,2,\cdots,K$$

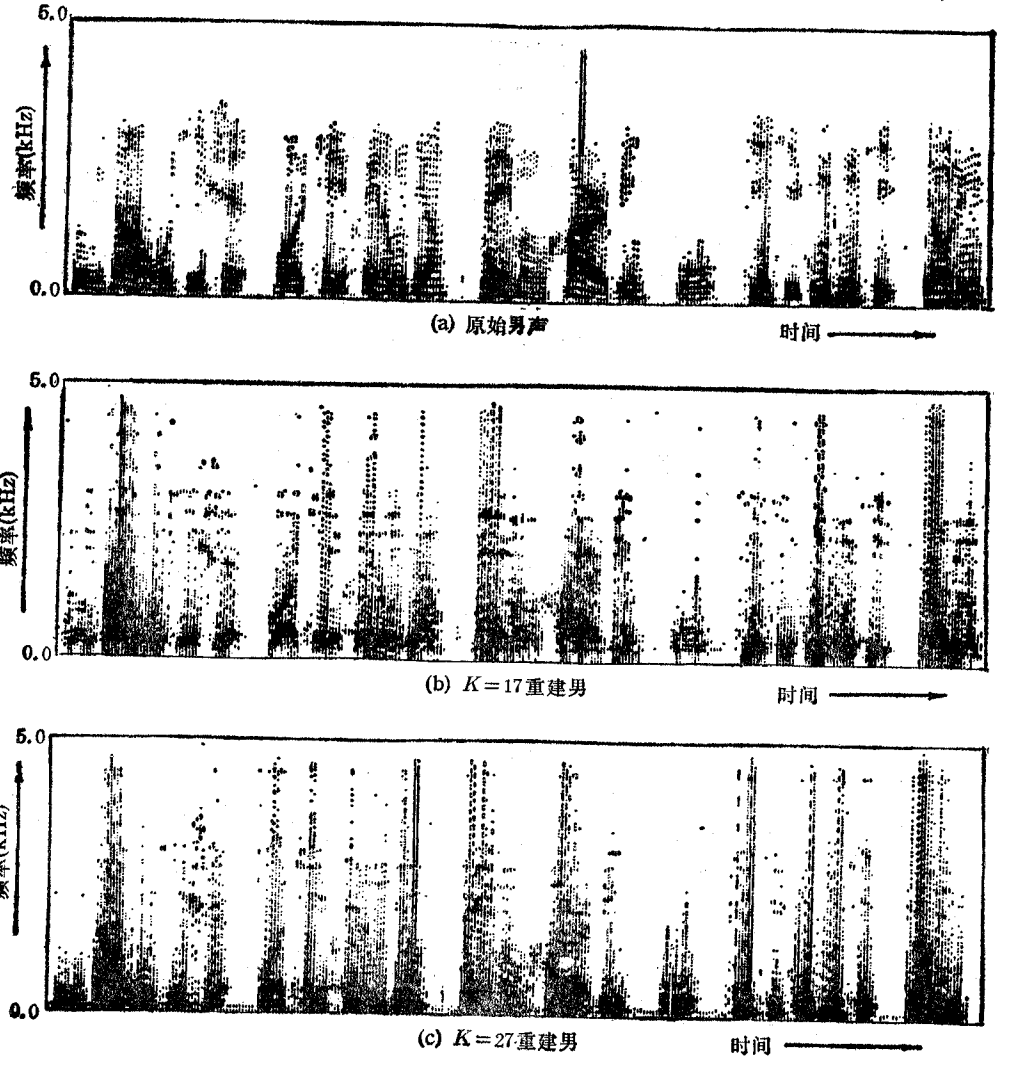


图 3 语谱图比较
(语谱图所表示的语声为"你到无锡市,我去黑龙江、内蒙、哈尔滨")
((b),(c) 应为重建男声)

其中 $\Omega(k)$ 为模拟频率 (rad/s)；然后将这 K 个余弦波进行简单相加，即当前帧重建语音为

$$s(n)=\sum_{k=1}^{k} A(k)\cos(\omega(k)\cdot n+\varphi(k)),\ n=1,2,\cdots,L_0$$

最后，对 $s(n)$ 进行能量平滑处理。

三、计算机模拟

我们用 FORTRAN 77 语言在 PC/AT 机上模拟了图 2 所示的语音分析/综合系统。输入语音数据库采用单个男声和单个女声，长度分别为 5.88 s 和 6.86 s。输入语音首先经过 4 kHz 的低通滤波，再以 10 kHz 速率进行采样。FFT 长度 (N) 取为 512 点，帧长取为 20 ms (即 $L_0=200$ 点)。听觉模型分别采用表 1 ($K=17$) 和表 2 ($K=27$) 两组参数。

计算机模拟结果表明：

(1) 尽管重建语音 (图 3 (b) 和 (c)) 伴有一些噪声，但与原始语音 (图 3 (a)) 比较起来，并未丢失重要的语音信息，证明了所建听觉模型虽然还不太完善，但其原理是基本正确的；

(2) 图 3 (c) 与图 3 (b) 相比更接近于图 3 (a)；听力测试结果也表明，图 3 (c) 的语音可懂度和自然度都较图 3 (b) 的有提高，特别是自然度有较大的改善；这说明我们对临界带宽参数进行的修正是有意义的，但是，这组修正参数还有待于进一步的完善。

四、总　　结

本文提出了一个新的听觉模型，该模型由滤波、检测和主频选取三部分组成。它们依次表征基底膜、内毛细胞和神经纤维的特性，因此该模型比较全面地反映了听觉系统特性。与目前仅有的几个听觉模型[3]比较起来，这是该模型的一个优点。基于所建立的听觉模型并结合修正的临界带宽参数，本文设计了一个语音分析/综合系统。实验表明 (图 3 (c))，重建语音是高可懂的、自然的，证明了本文听觉模型的正确性和可行性。当然，高质量新型语声处理系统的诞生还需要我们对听觉系统作进一步深入的研究。

参　考　文　献

[1] J. L. Flangan, Speech Analysis, Synthesis, and Perception, Academic Press, New York, (1965).

[2] 杨俊，樊昌信，听觉系统的生物物理模型，中国神经网络首届学术大会论文集，1990 年 12 月，北京，第171—174页。

[3] J. C. Anderson, Speech Analysis/Synthesis Based on Perception, TR-707, AD-A151 320, (1984).

[4] M. R. Schroeder, *Proc. IEEE*, 63(1975)9, 1332—1350.

[5] S. Seneff, Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model, ICASSP, 1984, San Diego, PP. 36.2.1—36.2.4.

[6] R. F. Lyon, Experiments with a Computational Model of the Cochlea, ICASSP, 1986, Japan, pp. 1975—1978.

[7] E. Zwicker, *J. Acoust. Soc. Am.*, 33(1961)2, 248—249.

[8] E. Zwicker et al., *J. Acoust. Soc. Am.*, 68(1980)5, 1523—1525.

[9] B. C. J. Moore, B. R. Glasberg, *J. Acoust. Soc. Am.*, 74(1983)3, 750—753.
[10] 杨俊，樊昌信，按听觉模型分析综合语音中频率匹配准则的改进，第四届语音图象通讯信号处理会议论文集，1989年10月，北京，第104—107页.

AUDITORY SYSTEM MODEL AND ITS APPLICATIONS

Yang Jun Fan Changxin
(*Xidian University, Xi'an* 710071)

Abstract A new auditory system model based on a combination of physiological and psychological acoustic data has been proposed. This model consists of a bank of nonuniform bandpass filters, detectors and main-frequency choosing mechanisms, they act as basilar membranes, inner hair cells and nerve fibers, respectively. Combining with the improved critical bandwidth parameters, the input to this model is analogous to the pressure at the eardrum, and the output of this model simulates various features of the firing patterns. The synthesizer obtains the resultant speech by use of the simple adding method. Computer simulations show that the resultant speech is highly intelligible and natural. The proposed model is correct, and the improvement of the critical bandwidth parameters is effective.

Key words Auditory system; Critical bandwidth; Speech analysis/synthesis