直接矢量量化方法在语音编码算法中的应用

赵群群,张雪英
ZHAO Qun-qun,ZHANG Xue-ying

太原理工大学 信息工程学院,太原 030024
College of Information Engineering,Taiyuan University of Technology,Taiyuan 030024,China
E-mail:zhao.qunqun@163.com

ZHAO Qun-qun,ZHANG Xue-ying.Application of direct vector quantization to speech coding algorithm.Computer Engineering and Applications,2008,44(15):39-42.

Abstract: The article introduces the principle of the Direct Vector Quantization (DVQ) algorithm which is applied to emulation encoder module and codebook search module in LD-CELP speech coding algorithm.The synthesis filter in emulation encoder module is replaces by the inverse perceptual filter,removing the operation of impulse response $h(n)$ in the codebook search module. The result show that the DVQ algorithm reduced its complexity and improved the efficiency of codebook search and the time is 3~5 second less than that of LD-CELP algorithm while keeping the quality.

Key words: vector quantization;codebook search;speech coding

摘 要:介绍了一种降低码书搜索复杂度的方法-直接矢量量化(DVQ)方法,将其应用于LD-CELP语音编码算法中的仿真译码器模块和码书搜索模块,用感觉加权逆滤波器代替仿真译码器模块中的综合滤波器,去除了码书搜索模块中冲激响应$h(n)$的运算。实验结果表明,利用直接矢量量化方法简化了码书搜索算法的复杂度,提高了码书搜索算法的效率,在运算时间方面比原始LD-CELP算法快3 s~5 s,同时保持了原编码算法合成语音的音质。

关键词:矢量量化;码书搜索;语音编码
DOI:10.3778/j.issn.1002-8331.2008.15.012 **文章编号:**1002-8331(2008)15-0039-04 **文献标识码:**A **中图分类号:**TN912.3

1 引言

1992年9月,国际电信联盟(ITU)(原国际电报电话咨询委员会(CCITT))公布了16 kb/s LD-CELP(Low-Delay Code Excited Linear Prediction)语音编码算法,即G.728标准。其特点是不直接从语音信号中提取语音短时谱与长时谱预测系数、增益因子等参数,而是利用50阶后向预测的方法得到。然而算法的编码复杂度相当大,仅编码过程就需要20 MIPS。人们一直希望能够减小算法复杂度,以满足更低延时的要求,而减小算法复杂度的主要途径之一是加速码书搜索过程[1]。本文在分析LD-CELP算法原理的基础上,利用文献[2]提出的直接矢量量化(Direct Vector Quantization(DVQ))思想,将其应用于G.728语音编码算法中。进一步阐述了感觉加权逆滤波器系数的选取、系数的更新过程、修改之后的码书搜索过程,以及增益码书量化的原理。结果表明在保证合成语音质量基本不变的前提下,减少了算法的计算量和复杂度,为语音编码器的实用化奠定了基础。

2 G.728编码算法原理

设$\boldsymbol{s}(n)$为待编码的语音矢量,其经过感觉加权滤波器$W(z)$的语音矢量为$\boldsymbol{v}(n)$($V(z)=S(z)W(z)$)。假定$\boldsymbol{u}(n)$为$\boldsymbol{v}(n)$的量化矢量,那么$\boldsymbol{u}(n)$经过感觉加权逆滤波器就得到最终的合成语音$\hat{s}(n)$。因为$\boldsymbol{v}(n)$,$\boldsymbol{u}(n)$是$\boldsymbol{s}(n)$,$\hat{s}(n)$经过滤波器后的形式,因此称$\boldsymbol{v}(n)$,$\boldsymbol{u}(n)$为感知域信号。量化误差信号$\boldsymbol{v}(n)-\boldsymbol{u}(n)$本质上应该是白噪声,但是由于感觉加权逆滤波器$1/W(z)$的存在,使得最终的量化误差信号$\boldsymbol{s}(n)-\hat{s}(n)$成为了有色频谱。选择合适的滤波器$W(z)$能够很大程度上增强编码器的频谱质量,因此这个感觉加权滤波器是编码算法中必不可少的一部分[2]。原理图如图1所示。

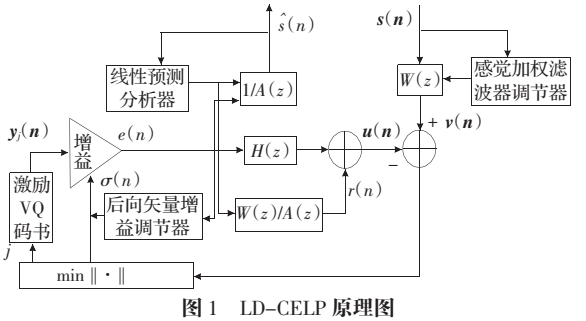


图1 LD-CELP原理图

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60472094);山西省自然科学基金(the Natural Science Foundation of Shanxi Province of China under Grant No.20051039)。

作者简介:赵群群(1981-),女,硕士研究生,主要研究方向为语音信号处理;张雪英(1964-),女,博士生导师,教授,主要研究方向为语音信号处理。

收稿日期:2007-09-06 **修回日期:**2007-11-30

激励信号 $e(n)$经滤波器 $H(z)=W(z)/A(z)$得到感知域输出 $\boldsymbol{u}(\boldsymbol{n})$,这里 $A(z)=\sum_{i=0}^{p}az^{-i}$,$p$ 为线性预测器的阶数。$e(n)=\sigma(n)\boldsymbol{y}_j(n)$,其中 $\sigma(n)$为后向自适应激励增益,$\boldsymbol{y}_j(n)$为第 j 个码矢量。$e(n)$经过 $1/A(z)$得到最终的输出信号$\hat{s}(n)$,这等效于 $\boldsymbol{u}(\boldsymbol{n})$经过感觉加权逆滤波器 $1/W(z)$就得到最终的合成语音 $\hat{s}(n)$。对于当前语音帧,在时间段$[0,\cdots,N-1]$,感知域的输出 $\boldsymbol{u}(\boldsymbol{n})$可以表示为:

$$\boldsymbol{u}(\boldsymbol{n})=r(n)+\sigma(n)\boldsymbol{y}_j(n)*h(n) \tag{1}$$

这里 $r(n)$是滤波器 $H(z)$的零输入响应,$h(n)$为 $H(z)$的冲激响应。假设输入矢量 $\boldsymbol{v}(\boldsymbol{n})$,$n=0,\cdots,N-1$,以及滤波器 $H(z)$,编码的主要任务就是找到最好的一组数$\{\sigma(n),y_j(n)\}$使得 $\|\boldsymbol{v}(\boldsymbol{n})-\boldsymbol{u}(\boldsymbol{n})\|$ 最小。为了做到这一点首先将激励优化为一个目标矢量:

$$x(n)=\boldsymbol{v}(\boldsymbol{n})-r(n) \tag{2}$$

然后找到:

$$\sigma(n)^*,j^*=\arg\min\|\boldsymbol{v}(\boldsymbol{n})-\boldsymbol{u}(\boldsymbol{n})\|=\arg\min_{\sigma(n),j}\|x(n)-\sigma(n)\boldsymbol{y}_j(n)*h(n)\| \tag{3}$$

滤波运算 $y_j(n)*h(n)$的次数与码书的大小一样,这样给编码过程带来了极大的运算量,也限制了码书的大小[2]。

3 直接矢量量化原理

直接矢量量化[2,3]的主要思想就是在感知域中避免卷积运算,为了达到这个目的,将激励用一个新的信号矢量 $\boldsymbol{t}_j(n)$来代替,表示为:

$$\boldsymbol{t}_j(n)=\boldsymbol{y}_j(n)*h(n)\quad 0\leqslant n\leqslant N-1 \tag{4}$$

这里 $\boldsymbol{t}_j(n)$代替 $\boldsymbol{y}_j(n)$成为一个新的固定码书。码书搜索过程变成一个简单的增益矢量量化问题:

$$\sigma(n)^*,j^*=\arg\min_{\sigma(n),j}\|x(n)-\sigma(n)\boldsymbol{t}_j(n)\| \tag{5}$$

在这个公式中没有涉及到滤波运算。在 DVQ 算法中,最终的感知域输出信号 $\boldsymbol{u}(\boldsymbol{n})$表示为:

$$\boldsymbol{u}(\boldsymbol{n})=r(n)+\sigma(n)^*\boldsymbol{t}_{j^*}(n)\quad 0\leqslant n\leqslant N-1 \tag{6}$$

一旦 $\boldsymbol{u}(\boldsymbol{n})$确定了,经过感觉加权逆滤波器就可以获得最终的量化语音$\hat{s}(n)(\hat{S}(z))$。

$$\hat{S}(z)=\frac{U(z)}{W(z)} \tag{7}$$

现采用不经过滤波的码本对信号直接矢量量化,即 $\boldsymbol{t}_j(n)$固定,这样码本搜索量大大减少。码本不经过滤波,意味着码本匹配不再考虑待编码信号自身特征,理论上 DVQ 得到的语音质量比传统的 VQ 方法要差些,但仍可得到较满意的语音质量。

4 直接矢量量化方法在 LD-CELP 算法中的应用

4.1 合成语音生成的具体算法

在直接矢量量化方法中,合成语音是加权语音矢量 $\boldsymbol{v}(\boldsymbol{n})$的量化矢量 $\boldsymbol{u}(\boldsymbol{n})$通过感觉加权逆滤波器 $1/W(z)$产生的。

在 G.728 中,感觉加权滤波器可以表示为:

$$W(z)=\frac{A_z(z)}{A_p(z)} \tag{8}$$

其中:

$$A_z(z)=1+\sum_{i=1}^{10}(q_i\gamma_1^i)z^{-i} \tag{9}$$

$$A_p(z)=1+\sum_{i=1}^{10}(q_i\gamma_2^i)z^{-i}\quad 0<\gamma_2<\gamma_1\leqslant 1 \tag{10}$$

这里 $\gamma_1=0.9$,$\gamma_2=0.6$。

在 DVQ-LD-CELP 中,感觉加权逆滤波器表示为:

$$\frac{1}{W(z)}=\frac{A_p(z)}{A_z(z)} \tag{11}$$

其中:

$$A_p(z)=1+\sum_{i=1}^{10}(q_i\gamma_p^i)z^{-i} \tag{12}$$

$$A_z(z)=1+\sum_{i=1}^{10}(q_i\gamma_z^i)z^{-i}\quad 0<\gamma_p<\gamma_z\leqslant 1 \tag{13}$$

为了确定 γ_p,γ_z的值,选择了 20 句汉语语句用于测试,其中包括 10 句男声,10 句女声。实验结果如表 1 所示。

表 1 γ_p,γ_z 不同值对应的平均分段信噪比

dB

γ_z	γ_p	平均分段信噪比
0.10	0.05	14.543 3
	0.06	14.647 1
	0.07	14.680 1
	0.08	14.724 2
	0.09	14.754 3
	0.10	14.752 0
0.11	0.10	14.741 6
0.12		14.745 8
0.13		14.705 7
0.14		14.623 4
0.15		14.529 5
0.20		14.013 7
0.30		12.574 1
0.40		11.017 9
0.50		9.396 6 5

另外,针对 $\gamma_p=0.09$,$\gamma_z=0.10$;$\gamma_p=0.10$,$\gamma_z=0.10$;$\gamma_p=0.10$,$\gamma_z=0.11$ 三组参数,选取了不同的语句进行进一步测试。其结果如表 2 所示。

表 2 $\gamma_p=0.09$,$\gamma_z=0.10$;$\gamma_p=0.10$,$\gamma_z=0.10$;$\gamma_p=0.10$,$\gamma_z=0.11$ 对应的平均分段信噪比

dB

SNR	参数取值		
	$\gamma_p=0.09,\gamma_z=0.10$	$\gamma_p=0.10,\gamma_z=0.10$	$\gamma_p=0.10,\gamma_z=0.11$
20 句	14.754 3	14.752 0	14.741 6
40 句	15.104 1	15.113 4	15.083 8
80 句	15.093 6	15.096 7	15.095 7
81 句女声	17.136 2	17.163 2	17.135 7
81 句男声	14.002 6	14.001 3	14.005 6

由表 2 可以看出,三组不同的参数取值,其平均分段信噪比相差并不悬殊,但 $\gamma_p=0.09$,$\gamma_z=0.10$ 的效果总体上稍好一些,因此在这里感觉加权逆滤波器的参数选定为 $\gamma_p=0.09$,$\gamma_z=0.10$。

这里,感觉加权逆滤波器的输入为:

$$\boldsymbol{u}(\boldsymbol{n})=r(n)+\boldsymbol{t}_j(n)\sigma(n) \tag{14}$$

其中,$r(n)$为零输入响应,$\boldsymbol{t}_j(n)$为重新训练好的码字,$\sigma(n)$为激

励增益。输出即为合成语音$\hat{s}(n)$。其原理图如图2所示。

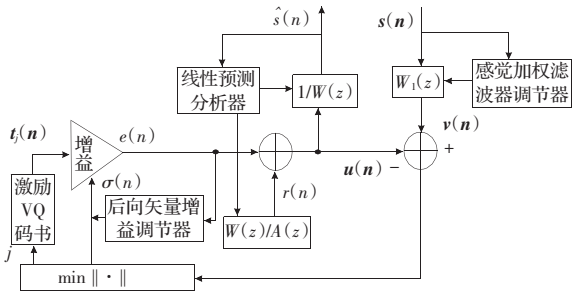


图2　DVQ-LD-CELP原理图

4.2　感觉加权逆滤波器系数的更新

在DVQ算法中，原始的G.728中感觉加权滤波器调节器已不能用于调节其逆滤波器，感觉加权逆滤波器的系数需要重新进行更新。主要是由于原始的G.728感觉加权滤波器调节器的输入为未量化的输入语音，而本算法中感觉加权逆滤波器是仿真译码器中的一部分，其系数调节需要利用量化(合成)语音进行。因此图2中$W_1(Z)$与$1/W(Z)$的系数是不同的。为了不额外增加算法的复杂度，本文利用综合滤波器的系数来更新感觉加权逆滤波器的系数，其原理如下。

综合滤波器的传递函数为：

$$F(z)=\frac{1}{A(z)}=\frac{1}{1+\sum_{i=1}^{50}a_i z^{-i}} \tag{15}$$

其中：

$$A(z)=1+\sum_{i=1}^{50}a_i z^{-i} \tag{16}$$

a_i为经过带宽扩展模块后得到的系数，其更新过程如图3所示。

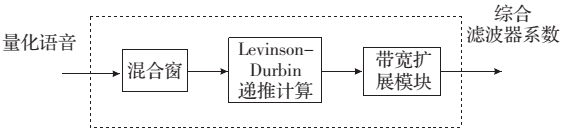


图3　综合滤波器系数更新

感觉加权逆滤波器的分子、分母分别为：

$$A_p(z)=1+\sum_{i=1}^{10}(q_i\gamma_p^i)z^{-i}=1+\sum_{i=1}^{10}a_{pi}z^{-i} \tag{17}$$

$$A_z(z)=1+\sum_{i=1}^{10}(q_i\gamma_z^i)z^{-i}=1+\sum_{i=1}^{10}a_{zi}z^{-i} \tag{18}$$

从式(16)、(17)、(18)可以看到，其滤波器的传递函数除了阶数、系数不一样，其他部分本质是相同的，因此本文将不经过带宽扩展模块所得到的系数设定为q_i，其分别乘以γ_p，γ_z得到a_{pi}和a_{zi}，系统在自适应周期的第4个语音矢量，即综合滤波器LPC系数计算的时候更新q_i，从而完成感觉加权逆滤波器的更新。

4.3　码书搜索过程

在DVQ算法中，码书仍然采用乘积码书的形式，即10 bit、1 024项的码书被分解为两个较小的码书：一个7 bit“波形码书”含有128个独立的码矢量和一个3 bit“增益码书”包含8个标量。针对128个独立的码矢量，本文用200句语音进行重新训练(其中包括100句男声，100句女声)。由于本算法去掉了$h(n)$的作用，因此码书搜索方法可以简化如下。

设经过激励增益的码矢量为：

$$\tilde{\boldsymbol{x}}_{ij}=\sigma(n)g_i y_j \tag{19}$$

码书搜索模块搜索下标i和j的最佳组合，这种组合使下面的均方误差最小：

$$D=\|x(n)-\tilde{\boldsymbol{x}}_i\|^2=\sigma^2(n)\|\hat{x}(n)-g_i y_j\|^2 \tag{20}$$

这里，$\hat{x}(n)=x(n)/\sigma(n)$为归一化目标矢量，展开上式：

$$D=\sigma^2(n)[\|\hat{x}(n)\|^2-2g_i\hat{x}(n)y_j+g_i^2\|y_j\|^2] \tag{21}$$

要使D最小，等价于使下面的$\hat{D}$最小：

$$\hat{D}=-2g_i\hat{x}(n)y_j+g_i^2E_j \tag{22}$$

这里$E_j=\|y_j\|^2$为128个波形码矢量的能量。

4.4　增益码书的量化

在对增益码书量化之前，首先计算出增益的精确表达形式，然后再对精确值进行量化得到增益量化值。

4.4.1　增益精确值求取过程

$\hat{D}$对g求导：

$$\hat{D}'=-2\hat{x}(n)y_j+2gE_j \tag{23}$$

令$\hat{D}'=0$，得：

$$g=\frac{\hat{x}(n)y_j}{E_j} \tag{24}$$

这时，$\hat{D}$最小，这就是g的精确表达式。对于增益的量化，本文采用如下叠代法来计算最佳量化特性。

4.4.2　确定最佳判决电平和最佳量化电平的原理

设信号$G(-\infty<G<+\infty)$的分布为$p(x)$，用$\hat{G}_i$表示G的量化值，误差记作：

$$\varepsilon(n)=\hat{G}_i-G \tag{25}$$

则量化噪声的方差为：

$$\sigma_e^2=E[\varepsilon^2(n)]=E[(\hat{G}_i-G)^2] \tag{26}$$

因为除去符号位后，对于$\hat{G}_i$只有k个量化离散值$\eta_i(i=0,1,2,\cdots,k-1)$，所以式(26)可改写为：

$$\sigma_e^2=\sum_{i=0}^{k-1}E[(\eta_i-G)^2]=\sum_{i=0}^{k-1}\int_{\xi_i}^{\xi_{i+1}}(\eta_i-x)^2p(x)\mathrm{d}x \tag{27}$$

本文希望选择参数$\{\xi_i\}(i=1,2,\cdots,k-1)$和$\{\eta_i\}(i=0,1,\cdots,k-1)$的集合，以使$\sigma_e^2$为最小，其中$\xi_0=-\infty$，$\xi_k=+\infty$，为此，令$\frac{\partial\sigma_e^2}{\partial\xi_i}=\frac{\partial\sigma_e^2}{\partial\eta_i}=0$。

因为$\frac{\partial\sigma_e^2}{\partial\eta_i}=0=2\eta_i\int_{\xi_i}^{\xi_{i+1}}p(x)\mathrm{d}x-2\int_{\xi_i}^{\xi_{i+1}}xp(x)\mathrm{d}x$，所以：

$$\eta_i=\frac{\int_{\xi_i}^{\xi_{i+1}}xp(x)\mathrm{d}x}{\int_{\xi_i}^{\xi_{i+1}}p(x)\mathrm{d}x}\quad(i=0,1,\cdots,k-1) \tag{28}$$

同理，因为$\frac{\partial\sigma_e^2}{\partial\xi_i}=(\eta_{i-1}^2-2\eta_{i-1}\xi_i+\xi_i^2)p(\xi_i)-(\eta_i^2-2\eta_i\xi_i+\xi_i^2)p(\xi_i)=0$，所以：

$$\xi_i=\frac{\eta_{i-1}+\eta_i}{2} \quad (i=1,2,\cdots,k-1) \tag{29}$$

式(28)表明,量化电平 η_i 最佳位置在 ξ_i 和 ξ_{i+1} 间隔内概率密度“矩心”上,式(29)表明,最佳判决电平 ξ_i 位于量化电平 η_{i-1} 和 η_i 的中间位置上。

这两组方程通常是非线性的,一般情况下,其求解必须用迭代法[4]。在$(-a,a)$范围内以 Δ 为步长,统计各 Δ 区间长度内的样点个数 f_i,除以总样点数 F,得到落入各小区间内的样点频率 f_i/F。以此频率 f_i/F 近似式(28)中的概率 $p(x)$。当 Δ 充分小的时候,可以认为在此 Δ 区间内信号服从均匀分布,运算中,以各 Δ 小区间的中点值表示所有落入该 Δ 区间的样点值,即得式(28)中的,给$\{\xi_i\}(i=1,2,\cdots,k-1)$和$\{\eta_i\}(i=0,1,\cdots,k-1)$赋上初值,使程序迭代计算新的$\{\xi_i\}(i=1,2,\cdots,k-1)$和$\{\eta_i\}(i=0,1,\cdots,k-1)$,直到其值稳定。

本文中,取 Δ=0.001,a 取不同值的时候对应的信噪比如表3所示。

表3 a 的不同取值对应的平均分段信噪比 dB

语音文件	a=3	a=5	a=6	a=8	原始
20句	13.782 1	14.754 3	14.113 0	13.848 8	14.806 5
40句	13.949 1	15.104 1	14.427 6	14.161 6	15.170 0
80句	13.951 2	15.093 6	14.388 1	14.107 4	15.071 5
81句女声	16.220 9	17.136 2	16.215 8	15.876 6	16.621 1
81句男声	12.762 8	14.002 6	13.306 2	13.062 0	14.124 9

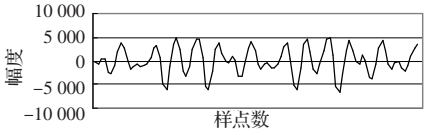
由表3可以看出,a=5时其语音的平均分段信噪比最好,a的取值范围决定了量化时所包含的增益真值的范围。若 a 选取的过小,那么包括的增益的真值比较少,不能够完全反应增益的特征;若 a 选取的过大,那么将有可能把噪音包括进来,反而降低了信噪比。这里取 a=5。

5 实验结果和结论

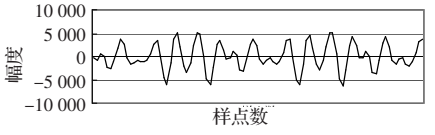
实验所用计算机为 Pentium® 4,CPU 2.93 GHz,384 MB内存。分别选择了20句(包括10句男声,10句女声),40句(包括20句男声,20句女声),80句(包括40句男声,40句女声),以及81句男声,81句女声汉语语句进行测试,其结果如表4所示。图4、图5又分别给出了一段男声、一段女声的输入语音和合成语音的对照图。

表4 DVQ-LD-CELP 与 LD-CELP 信噪比和时间对照表

语音文件	DVQ-LD-CELP		LD-CELP	
	SNR/dB	Time/s	SNR/dB	Time/s
20句	14.754 3	4.750 000	14.806 5	5.641 000
40句	15.104 1	9.625 000	15.170 0	12.110 000
80句	15.093 6	19.406 000	15.071 5	24.016 000
81句女声	17.136 2	21.563 000	16.621 1	27.218 000
81句男声	14.002 6	21.188 000	14.124 9	26.219 000

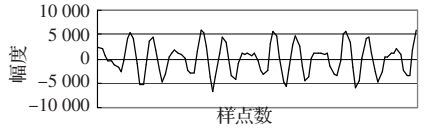


(a)男声输入语音

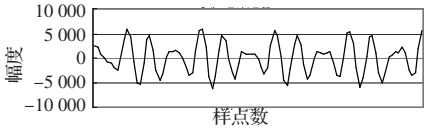


(b)男声合成语音

图4 男声输入语音与合成语音对照图



(a)女声输入语音



(b)女声合成语音

图5 女声输入语音与合成语音对照图

从表4可以看到,直接矢量量化 LD-CELP 与原始的 LD-CELP 在信噪比方面相差很小,平均只有 0.1 dB。在运算速度方面,直接矢量量化 LD-CELP 要明显比原始 LD-CELP 快,且语句越长越明显,80句的时候可以快 5 s。说明冲激响应矩阵在码书搜索过程中确实占有一定的运算量,去除 $h(n)$ 的作用对信噪比的影响不是很大,但却能提高运算的速度。从图4、图5可以看到,输入原始语音与合成语音在直观上基本相似。经主观非正式倾听,二者没有差别。

本文经过大量实验及理论分析,确定了感觉加权逆滤波器的系数 γ_p=0.09,γ_z=0.10,完成了其系数的更新过程。针对直接矢量量化算法,重新训练了新的码书。实验结果表明,DVQ 算法和原始的 728 算法在语音质量以及主观听觉方面相差不多,而 DVQ 算法明显比原始 728 算法的运算速度要快,对于算法的实时实现是很有利的。

参考文献:

[1] 张雪英.降低波形码书搜索复杂性的新方法[J].太原理工大学学报,1999,30(1):47-49.

[2] Shoham Y.On the use of direct vector quantization in lpc-based analysis by systhesis coding systems[C]//International Conference on Acoustics,Speech,and Signal Processing,Toronto Ont,Canada,1991.Washington,DC,USA:IEEE Computer Society,1991,1:5-8.

[3] 马霓,胡裕堂,韦岗.一种基于神经网络的 LD-CELP[J].深圳大学学报:理工版,1997,14(2-3):40-47.

[4] 沈江峰.8 kb/s 低延迟语音编码算法研究[D].太原:太原理工大学,2007.