一种 600bps 极低速率语音编码算法

丛 键 张知易
(现代通信国家重点实验室 成都 610041)

摘 要：该文针对抗干扰通信中对低速率语音编码算法的应用需求，提出了一种 600bps 极低速率语音编码算法，采用 6 帧超帧结构，超帧中包括 2 个基本帧与 4 个插值帧。插值帧的线性预测(LPC)参数采用基于闭环最优一阶线性预测的 4 阶段残差矩阵量化；在解码端，提出了闭环的激励脉冲幅度估计方法，提高了合成语音的自然度与鼻音音节的清晰度。该算法可以提供良好的合成语音质量，DRT 测试结果达到 88.55 分。

关键词：极低速率语音编码；多阶段矢量量化；多阶段矩阵量化
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2007)02-0429-05

A Very Low Bit Rate Speech Encoding Algorithm in 600bps

Cong Jian Zhang Zhi-yi
(*National Laboratory for Modern Communications*, *Chengdu* 610041, *China*)

Abstract: A very low bit rate speech encoding algorithm in 600bps is proposed in this paper for application in anti-jamming communication. Super-frame structure with 2 base frames and 4 interpolate frames is used, the LPC coefficients of interpolate frames is quantized with 4 stages residual matrix quantization based on optimal 1-order linear prediction. In the decoder, a closed loop estimation of the impulse magnitude is proposed to improve the naturalness of speech and the articulation of nasals. This speech coder achieves good quality of speech and the DRT score is 88.55。

Key words: Very low bit rate speech encoding; Multistage vector quantization; Multistage matrix quantization

1 引言

极低速率(<2400bps)语音编码在抗干扰通信方面有很强的应用需求，更低的数据传输速率意味着更好的抗信道干扰性能与更低的发射功率。基于参数编码的算法如 CELP 的多个版本、MELP[1,2] 与 MELPe、MBE 和 EMBE[3]、STC[4] 以及 WI[5,6] 系列等，可以在 2400bps 的编码速率上提供良好质量的合成语音。近年来，通过引入超帧提出了一些基于 MELP 与 MBE 的更低速率编码方案，如 AMBE2000 的 1200bps 算法和 MELP1200 算法[7] 等，而 NATO STANAG 4479 定义的 800bps 声码器采用 3 帧的超帧结构。

对于基于线性预测的参数编码算法，降低编码速率遇到最主要的瓶颈是线性预测(LPC)滤波器系数的量化编码，通常 LPC 系数会转换为线谱对(LSP)矢量进行量化，多阶段矢量量化(MSVQ)[8] 在每帧 21～25bit 的编码速率上可以实现很好的量化效果，而且多阶段的结构，使得这种量化方案在搜索算法的复杂度、码表存储资源以及抗误码方面具有更好的性能。在这个基础上，文献[9]中提出了一种残差多阶段矩阵量化(R-MSMQ)技术，并基于此实现了一种 1200bps 的语音编码算法[10]。

本文提出了一种 600bps 极低速率语音编码算法，算法采用 6 帧×25ms 的超帧结构。根据语音帧序列中清浊音特性的变化，将语音帧分为基本帧与插值帧。基本帧的 LPC 参数采用分类 4 阶段矢量量化；对于插值帧的 LPC 参数，则利用前后向的基本帧信息进行闭环最优一阶线性估计，估计残差矢量所构成的残差矩阵，采用 4 阶段的分类矩阵量化编码。6 帧语音信号的基音与清浊音标志采用分类的 2 阶段联合矢量量化；而能量参数则采用了 8bit 的分类矢量量化方案。

在解码端，为了解决由于语音合成滤波器记忆性所引起的信号能量变化与激励信号幅度变化的不一致，提出了闭环的激励脉冲幅度估计方法，提高了合成语音的自然度与鼻音音节的清晰度。最后为了评价本文 600bps 语音编码算法的性能，进行了单字清晰度(DRT)测试，测试结果达到了 88.55 分。

第 2 节中对整个编解码算法进行了描述，第 3 节提出了插值帧 LPC 参数的编码方案，第 4、5 节分别介绍了基音与能量参数的编解码处理过程。第 6 节是编码速率的分配方案，第 7 节则介绍了实验方案与 DRT 测试结果。

2 算法描述

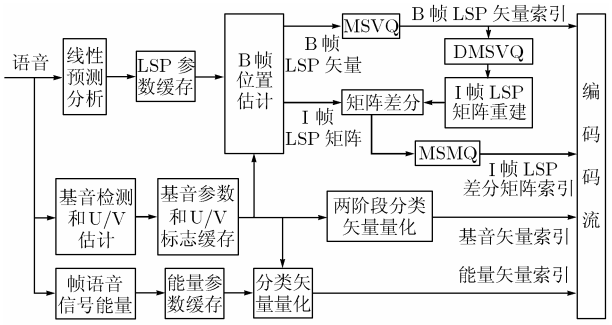
2400bps 成为标准速率之后，更低的标准速率一直没有确定，无论 1200bps 还是 800bps 都没有成为公认的标准速

2005-07-26 收到，2005-12-19 改回
现代通信国家重点实验室基金(51436050604DZ4003)资助课题

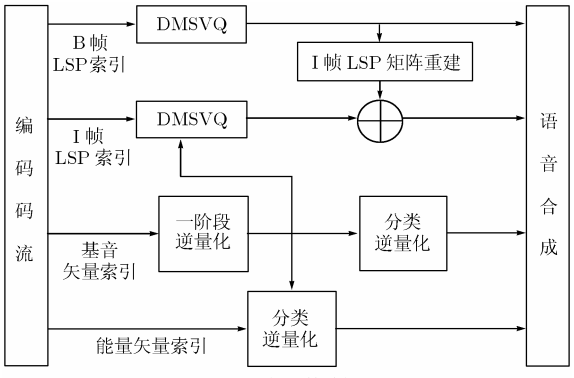
率。有希望成为标准的速率包括 600bps，300bps，后者被美军选为 MELP2400 之后的先进语音编码(Advanced Speech Encoding，ASE)技术的标准速率，并定义这种速率为超低比特率(ULBR)。因此我们的研究同时在这两种编码速率上展开，本文介绍速率 600bps 的编解码算法。

为了达到≤600bps 的编码速率，一是采用音节编码方案，而音节编码技术由于在语音自然度与音节识别所需存储资源等方面的缺陷，除了在 TTS(Text To Speech)系统中有良好应用外，在应用于实时语音通信方面存在瓶颈；另一途径就是在现有参数编码方案中利用帧间相关性，引入超帧(superframe)的结构，超帧方案对整个算法的结构有着重要的影响。为充分利用帧间冗余，本文采用 6 帧×25ms 的超帧，同 3 帧超帧相比，提高了编码效率与合成语音质量，但同时具有更大的编解码延迟与码表存储资源。

图 1(a)与图 1(b)给出了编解码算法的框图。编解码时，语音帧被分为基本帧和插值帧，在本文中分别记为 B 帧与 I 帧，同时浊音帧与清音帧分别记为 V 帧与 U 帧。每个超帧中包括 2 个 B 帧与 4 个 I 帧，B 帧的 LPC 参数采用失真更小的量化方案。由于人类听觉对浊音音节更敏感以及相邻 V 帧 LPC 参数具有很高的相关性，因此在 B 帧的选择中，V 帧特别是清浊音音节分界点上的 V 帧具有更高的优先级。B 帧的 LPC 参数采用 MSVQ，并且根据清浊音特性采用不同的量化码表，I 帧的 LPC 参数量化方案以及超帧的基音参数、能量参数的编解码方案将在后面介绍。



(a) 编码算法框图



(b) 解码算法框图

图 1

3 插值帧 LSP 矩阵的多阶段矩阵量化(MSMQ)

本文利用前后向 B 帧信息，基于一阶最优线性预测来估计 I 帧的 LSP 矢量。如果输入语音帧序列定义为 $f_i,\ i=1,2,\cdots$，并根据清浊音特性的变化分为 B 帧和 I 帧，分类时首先选择位于清浊音音节分界点上的 V 帧为 B 帧，如果分界点数量>2，则优先选择浊音音节起点位置的 V 帧为 B 帧，对于连续的清音帧或浊音帧，则选择 6 帧中的 2，5 帧为 B 帧。设 f_n, f_{n+m+1} 表示相邻的 B 帧，$f_{n+1},\cdots,f_{n+m}$ 为 I 帧，而 $\cdots \mathrm{lsp}_n, \mathrm{lsp}_{n+1},\cdots,\mathrm{lsp}_{n+m}, \mathrm{lsp}_{n+m+1},\cdots$ 为对应帧的 LSP 矢量，可由式(1)估计 I 帧的 LSP 矢量 $\hat{\mathrm{lsp}}_{n+k}[i]$：

$$\underset{a^i_{k,m+1-k},b^i_{k,m+1-k}}{\arg\min}\ E\left[\left(\mathrm{lsp}_{n+k}[i]-\hat{\mathrm{lsp}}_{n+k}[i]\right)^2\right] \tag{1a}$$

$$\hat{\mathrm{lsp}}_{n+k}[i]=a^i_{k,m+1-k}\cdot\mathrm{lsp}_n[i]+b^i_{k,m+1-k}\cdot\mathrm{lsp}_{n+m+1}[i] \tag{1b}$$

其中 $i=1,2,\cdots,L$，L 为线性预测分析的阶数，估计系数 $a^i_{k,m+1-k}$ 和 $b^i_{k,m+1-k}$ 可由式(2)得到

$$a^i_{k,m+1-k}=\left(r^i_0 r^i_k-r^i_{m+1-k}r^i_{m+1}\right)\Big/\left(r^{i^2}_0-r^{i^2}_{m+1}\right) \tag{2a}$$

$$b^i_{k,m+1-k}=\left(r^i_0 r^i_{m+1-k}-r^i_k r^i_{m+1}\right)\Big/\left(r^{i^2}_0-r^{i^2}_{m+1}\right) \tag{2b}$$

式(2)中 $r^i_k=E\left[\mathrm{lsp}_{n+k}[i]\mathrm{lsp}_n[i]\right]$；$r^i_k$，$a^i_{k,m+1-k}$ 和 $b^i_{k,m+1-k}$ 基于训练语音计算得到，表 1 中给出了部分估计系数。

残差矢量 $\Delta_{n+k}=\mathrm{lsp}_{n+k}-\hat{\mathrm{lsp}}_{n+k}$，对于超帧中 4 个I帧的LSP残差矢量所构成的$L\times4$矩阵，我们采用分类的 4 阶段矩阵量化，每层量化对应的比特开销为(7，7，6，6)。其中第 1 层为分类量化，设b1，b2 表示两个B帧在超帧中的位置，则由b2–b1(=2；=3；=4；=5)可将目标样本空间分为 4 类，并分别训练得到相应的码表，码表的训练采用了GLA(Generalized Lloyd Algorithm)算法[1]。本节中用于计算估计系数与训练矩阵量化码表采用了一段两小时的语音材料，并通过窗口滑动的方法产生容量 28 万的训练样本集，其中包括几十位男声、女声的话音；普通话和一些方言的话音；播音员与普通人的话音；没有背景噪声的话音、背景为加性高斯噪声的话音和室内噪声背景下的话音。

4 基音矢量的量化编码

对于 6 帧语音的基音参数和 U/V 标志，本文采取两阶段的分类矢量量化编码。首先根据超帧中 V 帧的数量，将样本空间分为 7 类，并基于超帧中 U/V 帧的分布特性将每类样本空间分为若干子类，为每一子类在第 1 层量化码表中分配固定数量的码矢如表 2，其中对于只包含 1 个 V 帧的超帧，由于浊音音节最少持续两个语音帧 50ms，对于孤立的浊音帧，在算法中判为清浊音检测错误，因此超帧中 V 帧只出现在超帧边界点上，对每种情况中 V 帧基音数值采用 7 层均匀量化，量化误差在第 2 层量化中采用 4bit 标量量化。

表 1　I 帧 LSP 矢量的一阶最优线性估计系数

i	1	2	3	4	5	6	7	8	9	10
$a_{1,2}^{i}$	0.65	0.66	0.73	0.60	0.70	0.76	0.70	0.64	0.61	0.63
$b_{1,2}^{i}$	0.35	0.34	0.27	0.40	0.30	0.24	0.30	0.36	0.39	0.37

表 2　第 1 层量化码表中码矢资源的分配

V 帧数量	0	1		2			3				4			5		6
子类码矢数量	1	7	7	24	24	24	40	40	40	40	45	45	45	67	67	90
				24	24	40	45	45	45	45	50	50	50			
码矢数量	1	14		160			340				285			134		90

第 1 层量化的速率为 10bit / 超帧，总的码矢数量为 1024；其次，针对所有的子类，分别产生训练样本集，并根据码矢的配额训练生成相应的子码表。第 2 层量化的编码速率为 4bit / 超帧，在编解码端，首先根据第 1 层量化与逆量化(码表索引查表)的结果，选择第 2 层量化使用的码表。对于全由 U 帧组成的超帧，不需要第 2 层基音量化，因此 4bit 作为保留；而只包含 1 个 V 帧的超帧，采用标量量化作为第 2 层量化；对于其他类，则分别产生训练样本，并为每一类生成 1 个 4bit 的第 2 层量化码表。

5　语音信号重建与能量参数编码

超帧的能量参数包括 6 帧语音信号能量组成的矢量，编码时信号能量由对数变换转换为分贝，并根据超帧中清浊音分布选择不同的量化码表，对能量参数进行分类矢量量化，量化采用无加权的欧氏距离。分类时，首先根据超帧中浊音帧的数量将能量矢量的样本空间分为 7 类，然后根据每类中清浊音帧在超帧中的时域分布，将每类样本空间进一步分为若干子类，总共 11 个子类；算法设计时，针对所有子类分别产生训练样本集，用 GLA 算法为每个子类生成 8bit 的码表，对能量参数编/解码时，首先根据超帧的清浊音分布选择所采用的子码表。

较早的基于线性预测分析的语音编码算法如 MELP2400，解码端语音合成时，激励脉冲的幅度在 1 帧内是固定不变的；改进的激励方案如EWI中，先确定每帧中心位置激励脉冲的幅度，然后通过线性插值或二次插值得到每个基音周期所对应的激励脉冲幅度。本文采用zinc函数[12]作为激励脉冲，zinc函数的定义如式(3)：

$$z(t) = K_1 \cdot \mathrm{sinc}(t) + K_2 \cdot \mathrm{cos\,c}(t) \tag{3a}$$

$$\mathrm{sinc}(t) = \sin\left(2\pi f_c(t)\right) \Big/ \left(2\pi f_c(t)\right),$$

$$\mathrm{cosc}(t) = \left(1 - \cos\left(2\pi f_c(t)\right)\right) \Big/ \left(2\pi f_c(t)\right) \tag{3b}$$

在解码端语音合成的过程中，由于合成滤波器的记忆性，因此合成信号能量幅度的变化规律会出现与激励脉冲幅度的变化规律不一致的情况，如图 2。特别当基音频率较高时更为明显。这种能量幅度变化的失真会影响合成语音的自然度，而且对于一些鼻音音节和浊辅音音节，信号能量幅度的变化对于音节的听觉分辨有重要的影响。因此本文提出了一种激励信号脉冲幅度的闭环估计算法，如图 3，使得输入语音信号的能量幅度变化规律可以在解码端语音合成时准确的重建。

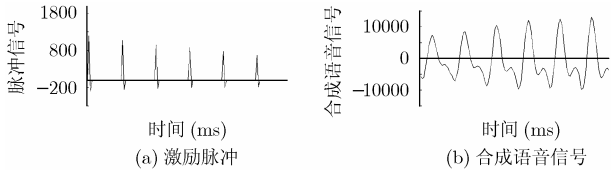


图 2　滤波器记忆性对合成语音的影响

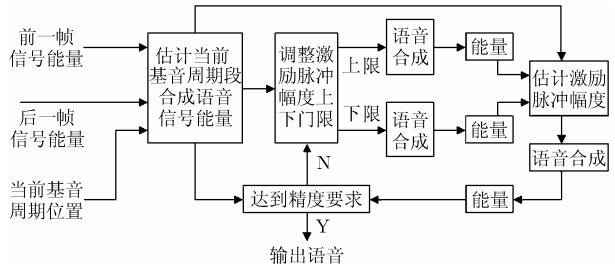


图 3　激励信号脉冲幅度的闭环估计方法

6　编码速率分配与码表存储资源

B帧的LSP参数采用 4 层矢量量化，清音码表与浊音码表每层量化的编码速率分配为 7＋6＋4＋4＝21bit，码表所需的存储资源为$(2^7+2^6+2^4+2^4)\times10\times2=(128+64+16+16)\times10\times2=4480$ 字＝8960 字节。I帧的LSP残差矩阵采用 4 层分类矩阵量化，每层量化的编码速率分配为 7＋7＋6＋6＝26bit，码表所需的存储资源为$(2^7\times4+2^7+2^6+2^6)\times40=(128\times4+128+64+64)\times40=30720$ 字＝61440 字节。

基音和U/V标志的联合矢量量化码表采用两层分类码表，编码速率分配为 10＋4＝14bit，码表所需的存储资源为$2^{10}\times6+2^4\times(2+3+4+5+6)=1024\times6+16\times20=6464$ 字节。6 帧语音信号的能量参数采用分类矢量量化，分类的数量为 11 类，编码速率分配为 8 bit，码表所需的存储资源为$2^8\times6\times11=256\times6\times11=16896$ 字节。

本文算法码表所需的存储资源为 8960＋61440＋6464＋16896＝93760 字节＝91.56 kB，编码延时为 6×25ms＝150ms。表 3 中给出了编码算法中超帧信号能量，基音和 LSP 参数的比特分配。

表 3 编码码流的比特分配

LSP 参数(6×25ms)		Pitch 与 U/V 标志(6×25ms)	信号能量(6×25ms)
B 帧	I 帧	10+4 bit	8 bit
21+21 bit	7+7+6+6 bit		
140+140 bps	173.33 bps	93.33 bps	53.33 bps
合计	90 bit / (6×25ms) =140×2+173.33+93.33+53.33 = 600 bps		

7 实验结果

为了评价本文提出的 600bps 极低速率语音编译码算法的性能，我们按照相关国家标准对算法进行了汉语清晰度诊断押韵测试(DRT)[13]。DRT 测试所需要的语音材料来源于根据国家标准 GB/T 16532－1996 建立的语音数据库，语音数据库中包括分别由(3 男＋3 女)6 个发音人采用标准清晰的普通话发音的 30 个单字发音字表，其中发音人的普通话能力达到国家标准 GB/T 13504－92(5.1.2，5.3～5.6)所规定的水平。语音材料的录制工作在录音工作室中完成，语音信号的记录采用 16kHz 采样、16bit 量化，并在计算机上转换为 8kHz 采样、16bit 量化的数字语音信号。

本文进行的 DRT 测试从语音数据库中选取了 2 男、2 女发音人，每人 3 张发音字表，(共 12 张)，经声码器编解码处理后供评听测试使用。考虑到测试的平衡性，根据国家标准 GB/T 16532，每个发音人的字表中都包含了分别从标准字表库 B1，B2 中选取的材料。每个发音字表包括 6 个测试项目——浊音性、鼻音性、送气性、低沉性、紧密性和持续性，每个测试项目分别包括 9 对押韵单字。

共有 10 个评听人员参与了 DRT 测试，实验资料的统计处理按照国家标准 GB/T 13504 进行，男声发音与女声发音的 6 个测试项目的 DRT 测试结果分别在表 4(a)和表 4(b)中给出，平均的 DRT 分数＝(89.36＋87.73)/2＝88.55。从测试结果可以看出，本文提出的语音编码算法处理男声发音的性能要略微好于女声，其中男声测试结果在鼻音性与持续性方面要好于女声，而女声发音则在低沉性的测试项目上明显好于男声。同时我们根据 SJ/T 20771-2000 对 600bps 算法进行了非正式的 MOS 测试，测试结果为 3.2 分。

8 结束语

美军在 2004 年开始了先进语音编码的研究工作，目的是为了制定 MELP 2400 和 MELPe 2400 之后的下一代语音编码标准，标准速率定为 300bps，项目计划到 2008 年完成。因此在设计 600bps 算法的同时，我们开展了 300bps 语音编码算法的研究，目前已经实现了 300bps 算法的编解码功能，DRT 测试达到 81.78 分。我们将来的研究重点将集中在提高 300bps 算法的性能与工程实现，使语音质量达到良好通话水平(DRT $>$ 85 分)。

表 4 (a)女声发音的 DRT 测试结果

测试项目	浊音性	鼻音性	送气性	低沉性	紧密性	持续性	总分
平均分	100.00	67.81	95.56	86.67	85.60	90.75	87.73
偏差	0.00	12.80	2.09	5.33	7.47	2.57	5.04

表 4 (b) 男声发音的 DRT 测试结果

测试项目	浊音性	鼻音性	送气性	低沉性	紧密性	持续性	总分
平均分	100.00	81.48	97.04	76.54	84.82	96.30	89.36
偏差	0.00	5.24	3.26	5.88	5.91	2.79	3.85

参 考 文 献

[1] McCree A V and Barnwell T P. A new mixed excitation LPC vocoder. In Proc. ICASSP, Toronto, Canada, 1991: 593–596.

[2] McCree A V and DeMartin J C. A 1.7kb/s MELP coder with improved analysis and quantization. In Proc. ICASSP, 1998, Seattle, vol.2: 593–596.

[3] Das A and Gersho A. Low rate multimode multiband spectral coding of speech. *International Journal of Speech Technology*, 1999, 2(4): 317–327.

[4] McAulay R J and Quatieri T F. Sinusoidal Coding. In Speech Coding Synthesis, Eds. The Netherlands: Elsevier, 1995, chap.4: 121–173.

[5] Kleijn W B. Speech coding below 4kb/s using waveform interpolation. In Proc. GLOBECOM, Phoenix, 1991, vol.3: 1879–1883.

[6] Shoham Y. Very low complexity interpolative speech coding at 1.2 to 2.4kb/s. In Proc. ICASSP, Munich, Germany, 1997: 1599–1602.

[7] Wang T, Koishida K, and Cuperman V, *et al.*. A 1200/2400 bps coding suite based on MELP. Proc. of IEEE Workshop on Speech Coding, Tsukuba, Japan, 2002: 6–9.

[8] LeBlanc W P, Bhattacharya B, and Mahmoud S A. Efficient search and design procedures for robust multistage vector quantization of LPC parameters for 4kbps speech coding. *IEEE Trans. on Speech and Audio Process.*, 1993, 1(4): 373–385.

[9] Nandkumar S, Swaminathan K, and Bhaskar U. Robust speech mode based LSF vector quantization for low bit rate speech coders. In Proc. ICASSP, Seattle, 1998: 41–44.

[10] Ozaydin S and Baykal B. A 1200 bps speech coder with LSP matrix quantization. IEEE Int. Conf. on ASSP, Salt Lake City, 2001: 677–680.

[11] Tsao C and Gray R M. Matrix quantizer design for LPC speech using the generalized Lloyd algorithm. *IEEE Trans. on ASSP*, 1985, ASSP-33(3): 537–545.

[12] Sukkar R A, LoCicero J L, and Picone J W. Decomposition of theLPC excitation using the sinc basis functions. *IEEE Trans. on ASSP*, 1989, 37(9): 1329–1341.

[13] 张知易, 王瑛. 几种中低速语音编码的音质评价实验. 第5届全国语音国家通信信号处理学术会议, 北京, 1997, 109–113.

丛　键：男，1973年生，工程师, 研究方向为语音编码、图像处理.

张知易：男，1942年生，研究员, 研究方向为语音编码、语音音质评价.