文章编号:1001 - 893X(2011)06 - 0056 - 05

基于二阶隐马尔可夫模型的清浊音恢复算法*

何洪华,徐敬德,计 哲,崔慧娟,唐 昆

(清华大学 电子工程系 清华信息科学与技术国家实验室,北京 100084)

摘 要:为了解决低速率语音编码中比特受限的问题,提出了一种基于二阶隐马尔可夫模型的清浊音参数恢复算法。该算法采用二阶隐马尔可夫模型,通过归一化的能量参数和 LPC 倒谱系数估计出序列中的全带清浊音判决和各个子带的清浊音度。解码器实现该算法后,编码器就无需对清浊音参数进行量化传输,从而节约了比特数。实验结果表明,该算法比基于 GMM 模型的算法能更好地恢复出清浊音信息,全带清浊音误判率减少了 5% ~ 20%,合成语音的 MOS 分比用 5 bit 的矢量量化(VQ)算法提高了 0.03 左右,达到了在节约比特数的同时也提高了语音质量的效果。

关键词:低速率语音编码;二阶隐马尔可夫模型;全带 V/U 判决;BPVC 恢复

中图分类号:TN912.32 **文献标识码**:A doi:10.3969/j.issn.1001 - 893x.2011.06.013

Voiced/Unvoiced Parameters Recovery Based on Second - Order Hidden Markov Model

HE Hong-hua, *XU Jing-de*, *JI Zhe*, *CUI Hui-juan*, *TANG Kun*

(Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: In order to solve the problem of limited number of bits in low bit rate speech coding, an algorithm using second - order Hidden Markov Model(HMM2) to recover the voiced/unvoiced parameters is proposed in this paper. The algorithm uses the normalized energy and linear prediction coding(LPC) coefficients to estimate the full-band V/U classification and the sub-band BPVC value. The algorithm can be implemented in the decoder, saving the bits originally used by V/U parameters and reducing the bit rate of speech coding. Experimental results show that the algorithm proposed can reduce the V/U classification error rate by 5% ~ 20% compared with the GMM algorithm, and improve the mean opinion score(MOS) of the synthesized speech signal by about 0.03 compared with the 5bit vector quantization(VQ), thereby greatly improves the estimation performance of the V/U parameters.

Key words: low-bit rate speech coding; second-order HMM; V/U classification; BPVC recovery

1 引 言

随着现代通信技术的不断进步,特别是光纤通信的发展使得通信的能力大幅提高。但是在信道价格昂贵的卫星通信、信道带宽极其有限的水声通信和信道环境恶劣的短波通信中,仍然存在着对超低速率的声码器的强烈需求。因此,有必要进一步研究300 bit/s甚至更低速率的语音编码器。

* **收稿日期**:2011 - 01 - 30;**修回日期**:2011 - 04 - 14
基金项目:国家自然科学基金资助项目(60572081)
Foundation Item: The National Natural Science Foundation of China(No.60572081)

在低速率语音参数编码算法中,一般在编码端对原始语音信号进行分析,提取各种能够表征语音信号的参数,如清浊音参数、线性预测系数(Linear Prediction Coding,LPC)、基音周期参数、能量参数等,对这些参数进行量化传输,然后在解码端使用反量化后的这些参数来合成语音信号[1]。因此,各个参数的量化性能直接影响着合成语音的质量。传统的低速率声码器通过衡量各个参数对语音质量的影响程度,然后给各个参数分配合理的比特数进行量化传输。然而在超低速率声码器中,分配给各个参数的比特数极其有限,各个参数的量化性能受到严重影响,从而影响了合成语音的质量。文献[2]提出了一种基于 GMM(Gaussian Mixture Models)模型的清浊音解码端恢复算法,使得浊音度参数无需传输,从而节约原本用于浊音度参数量化传输的比特。这样,节约出的比特数就可以分配给线性预测系数和基音周期等其它参数进行量化,使得其它参数的量化性能得到提高,从而使合成语音的整体性能也得到提高。但是文献[2]中的 GMM 模型忽略了语音信号参数具有时间相关性的事实。实际上,人的发音习惯相对稳定,相邻帧的清浊音参数之间相关性很大。为了更好地利用相邻帧的清浊音参数的相关性及其与能量参数、LPC 倒谱系数之间的统计相关性,本文提出了一种基于二阶隐马尔可夫模型的清浊音恢复算法。算法假定离散的清浊音为隐状态,归一化的能量参数和 LPC 倒谱系数组成的联合矢量为可观测状态,采用二阶隐马尔可夫模型估计出序列中的清浊音处于浊音状态的概率,将该值作为子带的清浊音模糊值。由于目前低速率声码器如 SELP[1] 和 MELP[3] 都是将语音信号按频率分为(0,0.5 kHz)、(0.5,1 kHz)、(1,2 kHz)、(2,3 kHz)、(3,4 kHz) 5 个子带,分别在各个子带内判断浊音度(BPVC),全带的 V/U 判决与第 1 子带的 BPVC 信息保持一致。因此,本文算法在恢复出各个子带的 BPVC 模糊值后,给第 1 子带的 BPVC 值设定一个门限即可以得到全带的 V/U 判决。

2　清浊音参数恢复算法

隐马尔可夫模型作为一种有效的语音信号统计模型,在语音识别和说话人识别研究中得到了广泛的应用[4-6]。本文假设每连续 N 个子帧组成一个超帧,超帧中的 BPVC 参数序列满足马尔可夫性,其中归一化能量参数和 LPC 倒谱系数(LPCC)为该马尔可夫链的可观测状态,BPVC 参数为隐状态,根据隐马尔可夫模型,通过归一化的能量参数和 LPC 系数来估计 BPVC 的状态。为了更好地利用 BPVC 参数的帧间相关性,算法采用二阶隐马尔可夫模型(HMM2)。

2.1　清浊音参数的 HMM2 模型

首先将带通浊音度的值分为两个状态,分别标记为 V(浊音)和 U(清音),当 BPVC 的值大于某个门限时就标记其为 V,否则为 U。假设 N 个子帧组成一个超帧,第 n 子帧第 b 个子带的清浊音参数的状态为 S_n^b,若 BPVC 状态满足二阶隐马尔可夫性,则有:

$$p(s_n^b \mid s_{n-1}^b, s_{n-2}^b) = p(s_n^b \mid s_{n-1}^b, s_{n-2}^b, \cdots, s_1^b) \quad (1)$$

不同语音帧的 BPVC 的状态关系可以由转移概率矩阵 $\boldsymbol{A}_1^b$、$\boldsymbol{A}_2^b$ 表示:

$$\begin{aligned} \boldsymbol{A}_1^b &= (a_{ij}^b)_{2\times 2} \\ \boldsymbol{A}_2^b &= (a_{ijk}^b)_{2\times 2\times 2} \\ a_{ij}^b &= p(S_n^b = j \mid S_{n-1}^b = i) \\ a_{ijk}^b &= p(S_n^b = k \mid S_{n-2}^b = i, S_{n-1}^b = j) \end{aligned} \quad (2)$$

式中,a_{ij}^b 表示第 $n-1$ 子帧第 b 子带的清浊音处于状态 i 时,第 n 个子帧的清浊音状态处于状态 j 的概率;a_{ijk}^b 表示第 $n-2$ 子帧的清浊音处于状态 i、第 $n-1$ 子帧的清浊音处于状态 j 时,第 n 子帧的清浊音处于状态 k 的概率;i 和 j 取值为 V 或 U,$b = 1,2,\cdots,5$,后面出现若无特别说明均取此值。

然后将十维的 LPC 系数转换成 12 维的 LPC 倒谱系数矢量 $\boldsymbol{l}$,并将其与归一化能量参数 $\bar{g}$ 组成一个联合矢量

$$\boldsymbol{z} = (\bar{g}, \boldsymbol{l}^{\mathrm{T}})^{\mathrm{T}}$$

式中,归一化能量参数 $\bar{g} = g/g_0$,g 为当前子帧的能量,g_0 为当前子帧的长时能量。当前子帧的长时能量的更新方式为 $g_0 = \alpha g + (1-\alpha) g_0$,$\alpha$ 为自适应修正的权重因子。

N 个子帧的联合矢量组成了马尔可夫链中的可观测序列,则 $p(z \mid S^b = i)$ 表示当第 b 个子带的清浊音参数处于状态 i 时出现观测矢量 z 的概率。在隐马尔可夫模型当中,观测矢量的概率密度通常由多个正态概率密度函数的线性叠加来逼近[7],即:

$$p(z \mid S^b = i) = \sum_{m=1}^{M} \boldsymbol{\alpha}_{i,m}^b N(z \mid \boldsymbol{\mu}_{i,m}^b, \boldsymbol{\Sigma}_{i,m}^b) \quad (3)$$

式中,$b = 1,2,3,4,5$,状态 i 取 U 或者 V,$\boldsymbol{\alpha}_{i,m}^b$、$\boldsymbol{\mu}_{i,m}^b$、$\boldsymbol{\Sigma}_{i,m}^b$ 分别表示第 m 个正态分布的权重、均值和协方

差矩阵,改变 $\alpha_{i,m}^b$、$\boldsymbol{\mu}_{i,m}^b$、$\boldsymbol{\Sigma}_{i,m}^b$就可以逼近各种实际情况中的概率密度函数。为了降低复杂度,文中假设 $\boldsymbol{\Sigma}_{i,m}^b$为对角阵,具体数值由 EM(Expectation-Maximization)算法训练得到。

2.2 清浊音参数恢复算法

根据上一节的假设,在已知上一超帧最后一子帧的清浊音状态和当前超帧各子帧观测矢量的条件下,算法采用 HMM2 模型通过以下动态规划过程估计当前子帧各子带的清浊音状态。

令前向概率 $\alpha^b(i,j,n)$表示第 $n-1$ 子帧第 b 子带的浊音度处于状态 i,第 n 子帧第 b 子带浊音度处于状态 j ,且观测矢量从第 1 帧到第 n 帧分别为 z_1 到 z_n 的概率,则有:

$$\alpha^b(i,j,n)=p(z_1,z_2,\cdots,z_n,S_{n-1}^b=i,S_n^b=j) \tag{4}$$

式中,$n=1,2,\cdots,N$。假设上一超帧的最后一子帧为第 0 帧,则初始化

$$\alpha^b(i,j,1)=I(S_0^b=i)a_{ij}^b p(z_1\mid S_1^b=j) \tag{5}$$

$I(S_0^b=i)$为逻辑函数,当 S_0^b 处于状态 i 时其值为 1,否则为 0。在得到初始值之后可以利用以下迭代公式可以求得 $\alpha^b(i,j,n)$:

$$\alpha^b(j,k,n)=\sum_{i=1}^{r}\alpha^b(i,j,n-1)a_{ijk}^b p(z_n\mid S_n^b=k) \tag{6}$$

式中,$n=2,3,\cdots,N$;$r=2$ 为状态数。

令后向概率 $\beta^b(i,j,n)$表示已知第子 $n-1$ 帧第 b 子带的浊音度状态为 i,第 n 子帧第 b 子带的浊音度状态为 j 的条件下,观测矢量从第 $n+1$ 子帧到第 N 子帧分别为 z_{n+1}到 z_N 的概率,则有:

$$\beta^b(i,j,n)=p(z_{n+1},z_{n+2,\cdots},z_N\mid S_{n-1}^b=i,S_n^b=j) \tag{7}$$

式中,$n=1,2,\cdots,N$。由初始条件 $\beta^b(i,j,N)=1$,通过以下公式迭代计算 $\beta^b(i,j,n)$:

$$\beta^b(i,j,n)=\sum_{k=1}^{r}a_{ijk}^b\beta^b(j,k,n+1)p(z_{n+1}\mid S_{n+1}^b=k) \tag{8}$$

式中,$n=2,3,\cdots,N$;$r=2$ 为状态数。

在通过动态规划迭代得到 $\alpha^b(i,j,n)$与 $\beta^b(i,j,n)$后,按下式可以计算出超帧中的第 n 子帧第 b 子带的浊音度状态分布:

$$p(S_n^b=j\mid S_0^b,z_1,\cdots,z_N)=\frac{\sum_{i=0}^{r}\alpha^b(i,j,n)\beta^b(i,j,n)}{\sum_{i=1}^{r}\sum_{j=1}^{r}\alpha^b(i,j,n)\beta^b(i,j,n)} \tag{9}$$

当 j 的状态为 V 时上式即为该语音帧的第 b 子带的 BPVC 参数处于状态 V 时的概率,该值即为第 n 子帧第 b 子带的模糊 BPVC 值。

一般认为全带清浊音判决与低子带的清浊音信息保持一致,因此可以根据第 1 子带的 BPVC 值直接判决全带的清浊音,如果下式成立则认为该语音帧为浊音,否则判决该语音帧为清音。

$$p(S_n^b=V\mid S_0^b,z_1,z_2,\cdots,z_n)>T_{uv},\text{令 } b=1 \tag{10}$$

式中,T_{uv}为预设的判决门限。这样通过本文算法既可以恢复出子带的 BPVC 模糊值,也可以得到全带的清浊音判决。

3 仿真实验

本文使用一段 114 min 的中文语音作训练库来训练状态转移矩阵和正态分布的各个参数,该数据库包含不同性别不同方言的说话人的不同语句。观测矢量逼近为 M 个正态分布的线性组合,一般来说,M 越大,逼近得越好,性能也就越好,但是复杂度也相应增加,而且随着 M 的增加,性能的提高会变得越来越不明显[2]。基于实际考虑选择 $M=8$ 进行模型训练。

3.1 清浊音参数恢复性能

为了去除 LSF 和能量的量化对恢复效果的影响,首先采用未量化的值来恢复 U/V 参数,计算算法对 U/V 参数的恢复效果,测试指标包括全带 V/U 判决的准确率以及 5 个子带的 BPVC 参数恢复误差,恢复误差的计算采用以下的加权失真:

$$E_{bp}(\hat{\boldsymbol{b}}_v,\boldsymbol{b}_v)=\sum_{i=1}^{K}w(b_v^i)(b_v^i-\hat{b}_v^i)^2 \tag{11}$$

式中,$\boldsymbol{b}_v=(b_v^1,b_v^2,\cdots,b_v^k)$为提取出的标准 BPVC 矢量[8];$\hat{\boldsymbol{b}}_v=(\hat{b}_v^1,\hat{b}_v^2,\cdots,\hat{b}_v^K)$为解码端恢复出来的 BPVC 矢量;$w(b_v^i)=\begin{cases}2^{5-i}w_v, & b_v^i=1\\ 2^{5-i}, & b_v^i=0\end{cases}$,$w_v=3$ 为对浊音的进一步加权,$i=1,2,\cdots,K$,K 为子带数 5。

全带 V/U 判决的测试语音采用带有全带 V/U 标注信息的 Keele 语音库,包括 10 个男女声说话人的话音内容,总时长为 5 min 36 s[2]。改变门限 T_{uv},得到清音误判为浊音的概率 P_{eu}和浊音误判为清音的概率 P_{ev}的相应变化曲线如图 1 所示。本文也实现了文献[2]的 GMM 算法,曲线越靠近左下方,错误率越低,性能越好。

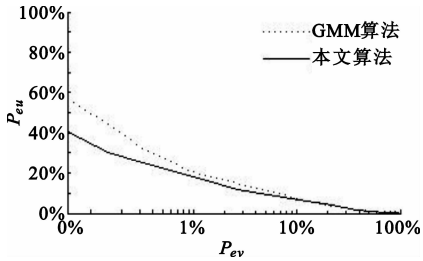

图 1　不同算法的 V/U 恢复性能曲线

Fig.1 V/U recovery performance of different algorithms

当 $P_{ev}\approx1\%$ 时,两种算法的 P_{eu} 如表 1 所示,由于浊音被误判对语音的影响远大于清音被误判的影响,所以通常在实际应用调整 T_{uv},使得 $P_{ev}<1\%$。由图 1 和表 1 可知,当 $P_{ev}<1\%$ 时,本文算法比 GMM 算法的 P_{eu} 减小了 5% ~ 20%,性能提高了 20% ~ 30%。

表 1　不同算法的 V/U 恢复性能

Table 1 V/U recovery performance of different algorithms

算法	P_{ev}/%	P_{eu}/%
GMM	1.00	22.94
HMM2	1.00	17.86

为了计算算法对各个子带 BPVC 参数的恢复性能,按式(11)计算 BPVC 参数的失真。表 2 列出了本文算法与 GMM 算法的失真结果,测试语句采用了 4 段长度约为 3 min 的训练集外标准测试语音。

表 2　不同算法的 BPVC 失真对比

Table 2 BPVC distortions of different algorithms

算法	语句 1	语句 2	语句 3	语句 4	平均
GMM	4.28	3.84	4.10	4.13	4.09
HMM2	3.81	3.26	3.58	3.61	3.57

由表 2 可知,本文算法相比 GMM 算法,BPVC 的失真减少了 12.7%。

由以上测试可知,本文算法相比于 GMM 算法能更好地恢复出全带 V/U 判决和子带 BPVC 参数。

3.2　对合成语音质量的影响

为了测试算法对整体语音性能的影响,在一种 SELP 声码器上进行了测试。该声码器以25 ms为一帧,采用 12 帧联合矢量量化的方式对 LSF 参数、能量参数(Gain)进行量化,由于在 SELP 声码器模型中,BPVC 要被用来辅助量化基音周期参数(Pitch),为了更客观地比较,基音周期参数采用直通方式,无量化失真,不同算法的各个参数的比特分配方式如表 3 所示。其中,VQ 算法采用 5 bit 对 BPVC 参数进行矢量量化(Vector Quantization)后传输;而 GMM 算法和 HMM2 算法不传输 BPVC 参数,只需在解码端根据量化后的 LSF 参数和能量参数分别采用 GMM 模型和 HMM2 模型对 V/U 参数进行恢复,并利用恢复的 V/U 参数对语音信号进行合成。

表 3　不同算法的比特分配方式

Table 3 Bits allocation of different algorithms

算法	LSF	Gain	BPVC	Pitch	总计
VQ	28	7	5	–	40
GMM	28	7	0	–	35
HMM2	28	7	0	–	35

测试语音采用 4 段长度约为 3 min 的训练集外标准测试语音。测试指标采用平均意见得分(Mean Opinion Score, MOS),测试过程采用国际电信联盟建议的 P.862 MOS 测试软件,对应于表 3 中不同的算法,相应的测试结果见表 4。

表 4　不同算法的语音 MOS 分对比

Table 4 Mean opinion score of different algorithms

算法	语句 1	语句 2	语句 3	语句 4	平均
VQ	2.558	2.478	2.528	2.525	2.522
GMM	2.576	2.516	2.546	2.562	2.550
HMM2	2.608	2.550	2.579	2.585	2.580

表 4 的测试结果表明,相较于 5 bit 的粗糙量化,采用 GMM 算法和本文算法后,客观 MOS 分都有不同程度的提高,且节省了 5 bit,而本文算法的 MOS 分比 GMM 算法又提高了 0.03,有效地提高了合成语音的质量。

4　结　论

在超低速率语音参数编码算法中,极其有限的比特数给各个参数的量化增加了困难。为此,本文提出了一种基于二阶隐马尔可夫模型的 BPVC 恢复算法,算法充分利用子带清浊音参数自身的时间相关性及其与能量、线性预测系数之间的统计相关性,采用二阶隐马尔可夫模型,用归一化能量参数和 LSF 参数来恢复 BPVC 参数,节省了原本用于 BPVC 参数量化传输的比特。实验结果表明,相比于 GMM 算法,本文算法能使全带清浊音误判率减少了 5% ~20%,使合成语音的平均 MOS 分提高了 0.03 左右。因此,在超低速率语音参数编码算法中,利用参数自身的时间相关性和各参数之间的统计相关性来进一步改善算法性能是下一步的研究方向。

参考文献：

[1] 李晔. 低速率语音编码技术与算法研究[D]. 北京：清华大学，2009.
LI Ye. Research on low bit rate speech coding techniques and algorithm[D]. Beijing: Tsinghua University, 2009. (in Chinese)

[2] Wei X, Dang X, Cui H, et al. Voiced/Unvoiced Classification Recovery in the Speech Decoder Based on GMM[C]// Proceedings of ICSP. Beijing: IEEE, 2008: 546 – 548.

[3] McCree V, Barnwell T. A mixed excitation LPC vocoder model for low bit rate speech coding[J]. IEEE Transactions on Speech Audio Processing, 1995, 3(4): 242 – 250.

[4] Rabiner L, Juang B H. Fundamentals of Speech Recognition [M]. New Jersey: Prentice – Hall, 1993: 321 – 386.

[5] Ismail Shahin. Using Second – Order Hidden Markov Model to Improve Speaker Identification Recognition Performance under Neutral Condition[C]//Proceedings of the 10th IEEE ICECS. Sharjah, United Arab Emirates: IEEE, 2003: 124 – 127.

[6] Jean – Francois Mari, Jean – Paul Haton, Abdelaziz kriouile. Automatic wordrecognition based on second – order hidden Markov models[J]. IEEE Transactions on Speech and Audio Processing, 1997(5): 22 – 25.

[7] 杨行峻，迟惠生. 语音信号数字处理[M]. 北京：电子工业出版社，1995：141 – 144.
YANG Xing – jun, CHI Hui – sheng. Voice digital signal processing[M]. Beijing: Publishing House of Electronics Industry, 1995: 141 – 144 (in Chinese)

[8] 李晔，洪侃，王童，等. 正弦激励线性预测声码器子带清浊音模糊判决[J]. 清华大学学报(自然科学版)，2008，48(7)：1101 – 1103.
LI Ye, HONG Kan. WANG Tong, et al. Fuzzy unvoiced/voiced decision – making for sub – bands for SELP vocoder [J]. Journal of Tsinghua University (Science & Technology Edition), 2008, 48(7): 1101 – 1103. (in Chinese)

作者简介：

何洪华(1985—)，男，湖南郴州人，2008 年获学士学位，现为硕士研究生，主要研究方向为低速率语音编码；

HE Hong – hua was born in Chenzhou, Hunan Province, in 1985. He received the B.S. degree in 2008. He is now a graduate student. His research direction is low – bit rate speech coding.

Email: hhonghua@gmail.com

徐敬德(1985—)，男，福建安南人，2007 年获学士学位，现为博士研究生，主要研究方向为低速率语音编码；

XU Jing – de was born in Annan, Fujian Province, in 1985. He received the B.S. degree in 2007. He is currently working toward the Ph.D. degree. His research direction is low – bit rate speech coding.

计　哲(1984—)，女，黑龙江大庆人，2006 年获学士学位，现为博士研究生，主要研究方向为低速率语音编码；

JI Zhe was born in Daqing, Heilongjiang Province, in 1984. She received the B.S. degree in 2006. She is currently working toward the Ph.D. degree. Her research direction is low – bit rate speech coding.

崔慧娟(1945—)，女，辽宁沈阳人，清华大学电子工程系教授，主要研究方向为信源编码、多媒体通信系统等；

CUI Hui – juan was born in Shenyang, Liaoning Province, in 1945. She is now a professor. Her research interests include signal source coding and multimedia communication system.

唐　昆(1945—)，男，江苏宜兴人，清华大学电子工程系教授，主要研究方向为数字通信、语音编码等领域。

TANG Kun was born in Yixing, Jiangsu Province, in 1945. He is now a professor. His research interests include communication, speech coding.