一种用于 WI 语音编码的相位预测式矢量量化方法

陈 悦　　鲍长春
(北京工业大学电子信息与控制工程学院　北京　100022)

摘　要：在传统的低比特率语音编码中，考虑到人耳对相位信息不敏感而经常忽略相位信息，这将导致语音粗糙、刺耳甚至音调发生改变。为了获得高质量的声码器，语音的相位信息是不能不考虑的。该文在散布相位矢量量化方法的基础上进一步去除了相位冗余，在波形内插(Waveform Interpolation，WI)编码模型中对相邻帧慢渐变波形(Slowly Evolving Waveform，SEW)的相位谱差值进行预测式矢量量化。实验发现，该方法大大改善了重建语音效果，明显提高了语音的自然度和清晰度。主观A/B测试结果显示，该方法与固定相位法相比，经 4~6 bit的相位量化可使合成语音质量得到显著的改善，相比散布相位矢量量化方法，女声的语音合成质量有所改进。

关键词：语音编码；波形内插；矢量量化
中图分类号：TN912.3　　**文献标识码**：A　　**文章编号**：1009-5896(2007)11-2672-04

A Predictive Phase Vector Quantization Method in WI Speech Coding

Chen Yue　　Bao Chang-chun
(*School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing* 100022, *China*)

Abstract: In traditional low bit-rate speech coding, considering that ears are not sensitive to phase information, the phase information is often neglected, and this will result in coarse and harsh speech quality, and it even may lead to inflection in pitch. In order to obtain a high-quality speech codec, the phase information of speech should be included in codec. In this paper, the phase redundancy is reduced further based on the dispersion phase vector quantization method. In the waveform interpolation (WI) speech coding model, the difference of SEW's phase spectra of conjoint frames is quantized using predictive vector quantization. The result of this scheme reveals that the speech quality is improved, and its naturalness and articulation are increased greatly. Subjective A/B listening test indicates that the reconstructed speech's quality of this method is better than that of fixed phase with 4–6 bit. Compared with the dispersion phase vector quantization method, the synthesis speech is slightly improved for female speakers.

Key words: Speech coding; Waveform interpolation; Vector quantization

1　引言

语音在人类通信中具有非常重要的地位。在现代通信中，语音通信作为人们交流信息的主要手段扮演着愈来愈重要的角色，而语音编码则是整个数字化通信网中最重要、最基本的组成部分之一。

近年来低速率语音编码算法得到了巨大的发展，人们越来越关注 4kbps及以下编码速率的高质量编码方案。波形内插语音编码是近十年来发展起来的一种性能优异的低速率语音编码方法，实现具有通信质量的低速率编码器已成为当今语音编码领域的一大研究热点。但目前基于WI[1-3]的低速率语音编码还没有达到理想的通信质量，其中一个主要原因就是相位信息的失真问题。

语音信号中相位信息的听觉感知效应长期被人们忽视，但随着对语音质量要求的提高，对该问题的深入研究已经显得越来越迫切了。由于人的听觉能感受到语音中任意频段的相位变化，因而为了提高语音的自然度，需要较为精确的重建原始语音的相位信息。

目前比较优秀的相位量化方法是Oded Gottesman提出的散布相位矢量量化方法[4–6]，它结合了感觉加权和分析合成(Analysis-by-Synthesis，AbS)方法，取得了较好的语音质量，在WI模型中每帧用 4~6 bit量化SEW的相位信息可以取得较好的效果。但是，该方案进行量化的对象——SEW相位谱仍存在冗余，因此，进一步降低量化比特数成为了可能。

本文的第 2 节简要介绍相位的听觉感知作用及冗余；第 3 节给出相位谱差值的预测式矢量量化及训练方案；第 4 节分析相位谱的重建方案；第 5 节给出基于该模型的实验结果；

2006-05-08 收到，2006-09-30 改回
国家自然科学基金(60372063)，北京市自然科学基金(4042009)和北京市教委科技发展计划项目(KM200710005001)资助课题

最后，得出全文的结论。

2　相位的听觉感知作用及冗余

WI算法编码原理如图 1 所示[7]。首先，语音信号经过线性预测分析得到 10 阶线性预测系数及线谱频率参数，本地解码的线谱频率参数变换为线性预测系数后构成线性预测滤波器，语音信号通过线性预测滤波器得到线性预测残差信号。然后，提取当前语音信号的基音周期。基音周期内插后根据基音周期值在残差信号域内提取特征波形。CW对齐后，将二维的CW表面通过低通滤波器滤波得到慢渐变波形(SEW)和快渐变波形(REW)。CW分解之后，利用SEW的功率分别归一化SEW和REW功率。归一化后的SEW，REW，SEW的功率以及基音，线谱频率(Line Spectrum Frequency, LSF)参数即为编码器输出参数。

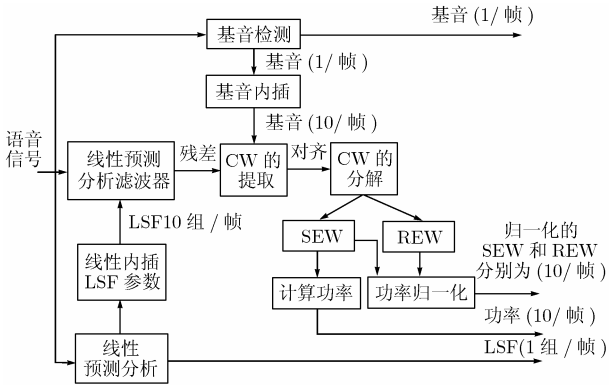


图 1　WI 算法编码原理

在语音中，相位是激励相位和声道相位的和。当激励相位在帧边界处线性变化、声道相位慢渐变时可以认为是相位一致的。保持相位的一致性在语音合成中有重要意义。合成的浊音语音如果相位不准确会造成回响，而清音则会出现嗡嗡声和冲击性的噪音。研究发现，声道相位在自然度方面对于语音并不重要，激励相位对于保持波形却起着重要的作用[8]。大量实验结果表明，保持语音信号的幅度谱不变，在其相位谱改变时只要重建信号在时域中的包络不变，重建语音和原始语音就不存在主观听觉上的差异[9]。因此，必须在浊音段保证激励相位的一致性。而对于清音段，尤其是在说话者基音频率较高的情况下，过度的相位周期性会导致断断续续的人工音调，为了消除这种现象需要强制激励正弦波形的相位不连贯[8]。WI编码将语音信号表示为一个具有二维特征波形(Characteristic Waveform，CW)表面的线性预测残差信号，因此经FIR滤波器分解得到的SEW和REW的相位信息对应语音的激励相位，且SEW具有浊音特征，REW具有清音特征。从编码有效性角度看，由于清音相位贡献小，可只传送浊音段相位以进一步降低比特率[10]。

Gottesman提出的低比特率波形内插编码的相位量化方法，以改进波形匹配为目的，对 SEW 的相位谱进行矢量量化，利用由幅度得到的感知加权函数来重建信号。为了进一步降低量化比特数，可以考虑对相位谱的其他表现形式进行量化。

首先，我们来分析一下相位谱中存在的冗余。图 2 为编码端CW对齐过程示例[11]。特征波形提取过程提供了每个CW的离散时间傅里叶级数(Discrete Time Fourier Series, DTFS)表示，然而这些CW一般是不同相的，即波形主特征在时间上没有对齐。通过对齐操作则可以得到CW的渐变描述。实质上，对齐过程也就是对当前CW进行循环时间移位，它等价于对DTFS系数加一个线性相位[12]。因此，对齐仅影响CW的相位谱，而不影响幅度谱。通过以上分析，可以发现对齐后的CW的相位谱是近似的，连续的相位谱间存在着冗余。

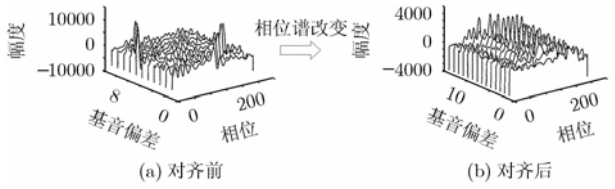


图 2　CW 对齐过程示例

由图 3 可以看出，对齐后每子帧间相位谱差值的动态范围相比原始相位谱的 $[-\pi,\pi]$ 大大减小。利用此性质，可以进一步改进相位的矢量量化方法。

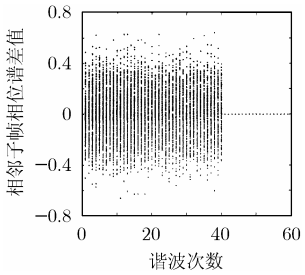


图 3　对齐后相邻子帧的相位谱差值

基于以上考虑，对由特征波形 CW 分解得到的类噪声的 REW 采用随机相位合成，对相邻帧 SEW 间的相位谱差值进行预测式矢量量化编码，从而减少相位量化所需比特数。

3　相位谱差值的量化及训练方案

相位谱携带一定语音的相关信息，SEW若用固定相位重建必然会影响重建语音质量。尽管对每次谐波的相位用 4 bit 均匀量化不会带来听觉上的感知差异[13]，但是其编码比特数过高。为了保证低比特率，本文参考文献[4]建立了SEW相位谱差值的预测式矢量量化模型，其原理框图如图 4 所示。图 4 中，$\boldsymbol{R}$和$\widehat{\boldsymbol{R}}$ 分别为SEW量化前后的DTFS，i 为码字序号，由上一帧SEW相位谱量化值和基音周期所对应的相位谱差值码书中的码字 $\hat{\varphi}_{\mathrm{err}_i}$ 计算得到 $\hat{\boldsymbol{\varphi}}$，$\boldsymbol{W}(z)$ 为感觉加权合成滤波器。量化的SEW幅度谱 $\left|\widehat{\boldsymbol{R}}\right|$ 与 $\hat{\boldsymbol{\varphi}}$ 结合得到 $\widehat{\boldsymbol{R}}$，通过使 $\boldsymbol{R}$ 与 $\widehat{\boldsymbol{R}}$ 之间的感觉加权失真测度 D_w 最小获得最佳的相位差值码字。$\boldsymbol{W}(z)$ 中的元素 $\boldsymbol{W}_{kk}$ 和失真 D_w 分别定义为[4]

$$W_{kk}=\left.\left|\frac{A(z/\gamma_1)}{A(z)A(z/\gamma_2)}\right|\right|_{z=e^{j(\frac{2\pi}{P})k}},0\le\gamma_2\le\gamma_1\le1 \quad (1)$$

$$\begin{aligned}D_w&=\frac{1}{K}\sum_{k=1}^{K}W_{kk}\left|R(k)-e^{j\hat{\varphi}(k)}\left|\widehat{R}(k)\right|\right|^2\\&=\frac{1}{K}\sum_{k=1}^{K}W_{kk}\left|R(k)-e^{j\left(\hat{\varphi}(k)_{\text{pre\_frame}}+10\hat{\varphi}_{\text{err}_i}(k)\right)}\left|\widehat{R}(k)\right|\right|^2\end{aligned} \quad (2)$$

式中 k 为谐波次数，P 为基音周期，$A(z)$ 为 LPC 多项式。

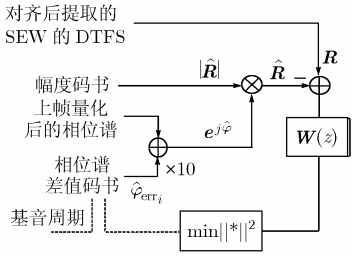


图 4 相位谱差值预测式矢量量化原理框图

在编码端首先需要计算相邻帧最后一子帧提取点处 SEW 相位谱之间的差值，满足失真最小的 $\hat{\varphi}_{\text{err}_i}$ 即为传输的相位谱差值矢量。相位谱量化方案如图 5 所示。

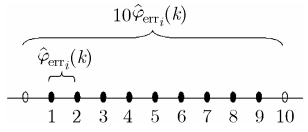


图 5 相位谱量化方案示意图

为了在有限的比特率下高效量化相位，将取值在[20,120]的基音周期根据语音的基频分布特点分为 8 个区间 $([20,30],(30,40],(40,50],(50,60],(60,70],(70,80],(80,100],(100,120])$ 分别设计最优码书[10]。类似于散布相位矢量量化的训练方法，为了保持相位变化的连续性，提取训练数据时使用重叠的训练聚类，训练的数据包括幅度谱、相位谱差值和感觉加权系数。平均全局失真测度公式为[4]

$$\bar{D}_{w,\text{global}}=\frac{1}{M}\sum_m\frac{1}{K_m}\sum_{k=1}^{K_m}W_{kk,m}\left|R(k)_m-e^{j\hat{\varphi}(k)_m}\left|\widehat{R}(k)\right|_m\right|^2 \quad (3)$$

其中 M 为训练数据个数，k 为谐波次数，K_m 为第 m 个训练数据的谐波阶数。质心应使平均全局失真测度最小，推导可知第 k 次谐波第 j 个聚类的质心满足

$$\tan\left(\hat{\varphi}(k)_{\text{err}_j}\right)=\frac{\sum_m\frac{1}{K_m}W_{kk,m}\left\{|R(k)|_m\left|\widehat{R}(k)\right|_m\sin\left[\varphi(k)_{\text{err}_m}\right]\right\}}{\sum_m\frac{1}{K_m}W_{kk,m}\left\{|R(k)|_m\left|\widehat{R}(k)\right|_m\cos\left[\varphi(k)_{\text{err}_m}\right]\right\}} \quad (4)$$

4 相位谱的重建方案

在解码端，首先根据基音周期找到相应的相位谱差值码书，然后解码得到当前帧与前一帧 SEW 相位谱之间的差值。然后，对上一帧量化后的相位谱依次加该相位谱差值得到每个 CW 提取点处的相位谱。最后，在帧内每个提取点处，将相位谱与解码并插值后的幅度谱合成，得到 SEW 的 DTFS 表示。具体框图如图 6 所示。

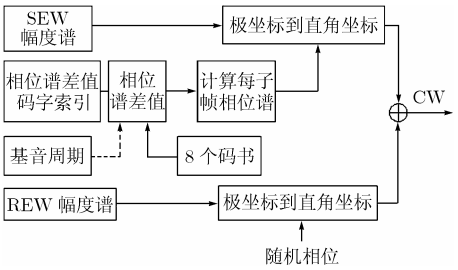


图 6 由相位谱差值重建 CW 框图

5 实验结果

训练数据采用 8k采样率的男女语音各 13 组，共 43670 帧，帧长 200 个样点。本文采用LBG算法进行相位谱差值码书的训练。训练过程中每组数据在 1~6 bit量化产生的平均失真如图 7 所示。对比发现，每组训练数据的平均失真大大小于散布相位矢量量化方案[10]。

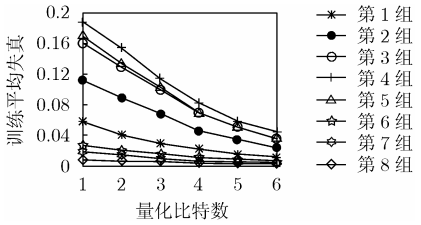
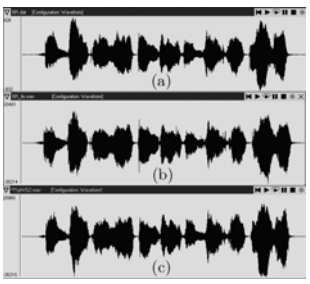


图 7 相位码书在不同比特下的平均失真

实验发现用 4~6 bit 量化的效果已经比较令人满意，大量主观测试结果表明其自然度和清晰度明显优于采用固定相位法的合成语音。将此方案应用到 WI 编码模型中(仅量化 SEW 相位信息)，以一段女声为例，其量化前后波形对比如图 8 所示。



(a)原始语音 (b)固定相位合成的语音
(c) 6bit 相位差值量化合成的语音
图 8 相位量化前后语音波形对比

本文使用ITU-T 的语音质量感觉评估(Perceptual Evalua-

tion of Speech Quality, PESQ)作为对比评价，和散布相位矢量量化算法比较的结果如表 1 所示。

表 1　合成语音效果对比(PESQ 分)

	6 bit 散布相位矢量量化	6 bit 相位谱预测式矢量量化
女声	2.839	2.898
男声	3.091	3.054

6　结束语

本文分析了相位的听觉感知作用和冗余的存在，并给出了相位谱差值的预测式矢量量化及码书训练方案。该方案首先计算当前帧与前一帧 SEW 的相位谱差值，然后寻找基音周期对应码书中令失真测度最小的相位谱差值码字作为量化值。实验结果显示，通过有效的矢量量化，合成语音的自然度和清晰度相比用固定相位合成 SEW 有了显著提高；相比散布相位矢量量化方案，本方案进行训练时收敛速度明显加快，同时女声的语音合成质量有所改进。

参 考 文 献

[1] Kleijn W B and Haagen J. Waveform Interpolation. Speech Coding and Synthesis, Amsterdam: Elsevier Science B. V., Chapter 5, 1995: 175–207.

[2] Kleijn W B and Haagen J. A speech coder based on decomposition of characteristic waveforms. IEEE ICASSP, Detroit, USA, 1995, vol.1: 508–511.

[3] Kleijn W B and Haagen J. Transformation and decomposition of the speech signal for coding. *IEEE Signal Processing Letters*, 1994, 1(9): 136–139.

[4] Gottesman O. Dispersion phase vector quantization for enhancement of waveform interpolative coder. IEEE ICASSP, Phoenix, Arizona, USA ,1999, vol.1: 269–272.

[5] Gottesman O and Gersho A. Enhanced waveform interpolative coding at low bit-rate. *IEEE Trans. on Speech and Audio Sig -nal Processing*, 2001, 9(8): 786–798.

[6] Gottesman O and Gersho A. Enhanced analysis-by-synthesis waveform interpolative coding at 4kbps. [Ph. D Dissertation], University of California. 1999: 1443–1446.

[7] 朱娜娜. 2kbps 波形内插语音编码算法的研究. [硕士论文]，北京工业大学. 2003: 10–70.

[8] Quatieri T F and McAulay R J. Phase coherence in speech recons -truction for enhancement and coding applications. IEEE IC -ASSP, Glasgow, Scotland, 1989, vol.1: 207–210.

[9] 同鸣等. 语音信号中相位信息的听觉感知研究. 西安交通大学学报，2003, 37(12): 1288–1291.

[10] 陈悦，鲍长春. WI 语音编码中相位信息的量化与重建. 信号处理, 2005, 21(4A): 164–167.

[11] Chong-White N R and Burnett I S. Accurate, critically sampled characteristic waveform surface construction for waveform interpolation decomposition. *IEE Electronics Letters*, 2000, 36(14): 1245–1247.

[12] 鲍长春. 低比特率数字语音编码基础[M]. 北京：北京工业大学出版社. 2001: 233–257.

[13] Kim Doh-Suk and Kim Moo Young. On the perceptual weighting function for phase quantization of speech. IEEE Workshop on Speech Coding, Wisconsin, USA, 2000: 62–64.

陈　悦：　女，1981 年生，硕士，从事语音信号处理及相位重建问题的研究.

鲍长春：　男，1965 年生，博士，教授，博士生导师，国际语音通信学会(ISCA)会员，中国电子学会理事，信号处理学会委员，《通信学报》与《信号处理学报》编委，主要研究领域为数字信号处理与语音编码.