# 基于信号规整和稀疏变换的语音与音频分层编码方法

李晓明,鲍长春,贾懋▮
(北京工业大学电子信息与控制工程学院语音与音频信号处理研究室,北京 100124)

**摘 要:** 基于语音和音频信号的固有周期性特征,本文构建了一种适合语音和音频信号的统一分析/合成模型,并分别在 24kbps 和 32kbps 码率下,实现了对宽带语音和音频信号的高质量分层编码.首先,本文将具有时变周期的输入信号规整为具有固定周期的信号,并对规整后的周期信号构建规整矩阵;其次,对规整矩阵的行和列分别进行调制叠接变换(MLT)和离散余弦变换(DCT),完成规整矩阵的稀疏化;最后,利用分带量化和矢量哈夫曼编码完成稀疏矩阵元素的量化和编码.主客观测试结果表明,本文所提方法的语音、音频及其混合信号的编码质量均优于同等速率下的 ITU-T G.722.1 和 AMR-WB 编码器.

**关键词:** 语音编码; 音频编码; 信号规整; 稀疏变换
**中图分类号:** TN912.3 **文献标识码:** A **文章编号:** 0372-2112 (2015)07-1286-08
**电子学报 URL:** http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.07.006

# The Layered Coding of Speech and Audio Signals Based on Signal Warp and Sparse Transform

LI Xiao-ming, BAO Chang-chun, JIA Mao-shen
(*Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing* 100124, *China*)

**Abstract:** Based on the periodic characteristics of speech and audio, a layered coding method by using uniform analysis and synthesis model is proposed in this paper. The constructed coder can perform equally well on speech and audio at the bit rates of 24kbps and 32kbps. First, the input signal which has time-varying period is warped into a constant period signal. Second, a sparse representation of the warped signal is achieved by applying the MLT and DCT on the warped matrix derived from the warped signal. Finally, the sub-band quantization and Huffman coding are applied on the transform coefficients. Both the objective PESQ/PEAQ results and the subjective A/B listening tests show that the proposed coder outperforms the ITU-T G.722.1 and AMR-WB codec.

**Key words:** speech coding; audio coding; signal warping; sparse transform

## 1 引言

在保持一定听觉质量的前期下,现存语音编码[1,2]和音频编码[3]都是通过去除信号的冗余信息来实现压缩编码.语音编码[4,5]基于人类语音的产生模型,通过去除信号远样点间和近样点间的相关性,实现语音信号的高质量编码,如:码激励线性预测语音编码[6,7].而音频编码则是借助心理声学模型,通过时-频变换[8,9]和听觉掩蔽机理[10]去除信号间的相关性,实现对音频信号的编码,如:MPEG MP3[11]音频编解码器.语音和音频信号的建模差异,使得它们无法同时对语音和音频信号进行有效编码[12],导致实际应用不够便捷和灵活.

鉴于上述问题,MPEG 音频标准制定组提出了一种基于信号类型判别的通用语音与音频编码(Unified Speech and Audio Coding, USAC)方法[12].通过对语音和音频信号分别采用 AMR-WB +[7] 和 MPEG HE-AAC[13] 编码,实现对语音和音频信号的通用编码.但其编码质量过分依赖输入信号类型的准确判决,对语音和音频混合

收稿日期:2014-01-03;修回日期:2014-03-17;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61072089,No.61201197);北京市教委科技计划面上项目(No.KM201310005008);教育部博士学科点专项科研基金新教师基金(No.20121103120017);北京工业大学第 12 届研究生科技基金(No.ykj-2013-9563)

信号处理能力较弱,导致整体性能不够理想.

针对上述问题,本文通过发掘输入语音和音频信号的固有谐波特征,利用统一分析/合成模型实现对语音和音频信号的通用编码.首先,本文将输入信号规整为具有恒定周期的信号,以去除输入信号周期的时变特征;其次,利用规整信号相邻周期信号波形的相似性,通过 MLT[8]和 DCT[9,10]对相邻若干周期规整信号进行稀疏变换,将变换系数的能量集中于少数频带;最后,对稀疏的变换系数进行量化编码.

## 2　信号规整

语音和音频信号的周期性特征使得相邻周期间信号具有极强的相关性,这为去除信号冗余信息,实现稀疏信息的编码提供了可能[14].由于信号周期和信号本身是时变的,周期波形的长度和形状存在较大差异,不利于去除周期波形间的相关性.为此,本文将输入信号规整为具有恒定周期的信号,以去除输入信号的时变周期.

### 2.1　信号规整的基本思想

信号规整[15]的基本思想是:对时间依赖的输入信号的周期及信号本身进行内插,得到如图 1 所示的具有恒定周期的规整信号.

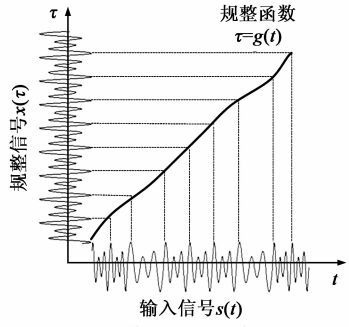



图1　信号规整示意图

输入信号 $s(t)$可表示为如下傅立叶级数形式:

$$s(t) = \sum_{k\in Z}(a_k\cos(k\varphi(t)) + b_k\sin(k\varphi(t))) \quad (1)$$

式中 $a_k$ 和 $b_k$ 为傅立叶级数的系数,$Z$ 为傅立叶级数展开的项数,$\varphi(t)$为与输入信号瞬时基频 $f_0(t)$相对应的瞬时相位,即有如下关系:

$$f_0(t) = \frac{1}{p(t)} = \frac{1}{2\pi}\frac{\mathrm{d}\varphi(t)}{\mathrm{d}t} \quad (2)$$

式中,$p(t)$为输入信号在 $t$ 时刻的周期.

本文通过规整函数来描述输入信号与规整信号间的关系:

$$\tau = g(t) = \frac{1}{F_0}\int_0^t f_0(s)\mathrm{d}s \quad (3)$$

式中,$\tau$ 为规整信号的时间变量,$F_0$ 为规整信号的固定基频,结合式(2)可得:

$$\varphi(t) = 2\pi\int_0^t f_0(s)\mathrm{d}s = 2\pi F_0 g(t) \quad (4)$$

将式(4)代入式(1),可得:

$$\begin{aligned} s(t) &= \sum_{k\in Z}(a_k\cos(k2\pi F_0 g(t)) + b_k\sin(k2\pi F_0 g(t))) \\ &= \sum_{k\in Z}(a_k\cos(k2\pi F_0\tau) + b_k\sin(k2\pi F_0\tau)) \\ &= x(\tau) \end{aligned} \quad (5)$$

式中 $x(\tau)$为规整信号,$\tau$ 为规整信号的连续时间变量.

此时,规整信号 $x(\tau)$的周期 $P = 1/F_0$ 在任意时刻 $\tau$ 均为恒定的常数,实现了对具有时变周期输入信号的规整.为保证不丢失原始信号的频率成分,规整信号的周期必须满足:$P \geqslant P_{\max}$,即规整信号的周期要大于等于原始信号的最大周期 $p_{\max}$.

### 2.2　信号规整的原理

对于实际语音和音频信号,本文采用分段插值的方式来实现对输入信号的规整,图 2 给出了信号规整的原理框图.

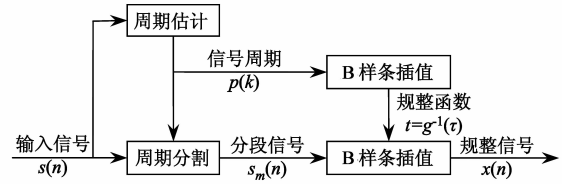



图2　信号规整原理框图

由于无法准确获得输入信号瞬时周期的数学表示,故本文利用信号的周期信息,通过 B-样条插值得到规整函数.同时,对于离散的输入信号 $s(n)$,规整前后信号的样本点间并非整数对应关系,势必会用到输入信号采样点间的数值,因此,需再次利用 B-样条插值,对输入信号进行展开,以得到离散的规整信号 $x(n)$,具体算法如下:

首先,利用自相关检测算法[1]对输入信号 $s(n)$进行周期估计,得到输入信号第 $k$ 帧的周期 $p(k)$.为准确描述输入信号周期的时变性,本文对相邻两帧信号的周期进行线性插值,得到信号的时变周期 $p(k,m)$:

$$p(k,m) = \frac{M-m}{M-1}p(k) + \frac{m-1}{M-1}p(k+1) \quad (6)$$

式中,$k$ 为当前分析帧索引,$m\in[1,M]$为周期波形索引,其中 $M = \lfloor N/p(k)\rfloor$为当前帧包含的周期波形的个数,$N$ 为分析帧长.

其次,用 $p(k,m)$将当前帧分割成 $M$ 段信号 $s_m(n)$,未包含在 $M$ 段信号中的样点留作下一帧分割用.进而对 $s_m(n)$的起止时间进行三次 B-样条插值[16],以构建函数 $t = g^{-1}(\tau)$:

$$t = g^{-1}(\tau) = \sum_{m\in M} z_m\beta^3(\tau - mP) \quad (7)$$

式中，$z_m$ 为三次 B-样条插值函数在 $s_m(n)$ 终止时间处的控制节点[17]，$\beta^3(\cdot)$ 为三次 B-样条函数，$P$ 为规整信号的周期.通过插值，与 $s_m(n)$ 相对应的规整信号的终止时间被拉伸至 $m\cdot P$ 处，使得规整信号具有相同的周期 $P$.

最后，基于函数 $t=g^{-1}(\tau)$，对当前帧的各段信号进行三次 B-样条插值，计算规整信号在 $\tau=nT_0$ 处的幅值，其中 $T_0$ 为输入信号的采样间隔，最终得到离散的规整信号 $x(n)$：

$$\begin{aligned} x(n) &= x(\tau)|_{\tau=nT_0} \\ &= s(g^{-1}(\tau))|_{\tau=nT_0} \\ &= \sum_{l\in D} c_l \beta^3\Big(\sum_{m\in M} z_m \beta^3(nT_0-mP)-lT_0\Big) \end{aligned} \tag{8}$$

式中，$c_l$ 为输入信号第 $l$ 个采样点的 B-样条插值控制节点，$D$ 为第 $m$ 段输入信号的样点数，$\beta^3(\cdot)$ 为三次 B-样条函数，$T_0$ 为输入信号的采样间隔.图 3 给出了对连续两帧语音信号进行规整的结果.

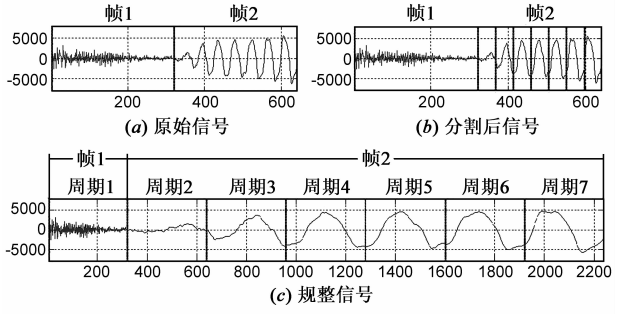



**图3 实际语音信号的周期分割与规整结果**

图 3($a$)为包含清音和浊音的待规整语音信号.首先，由于清音不具有周期性，故本文将清音的周期设定为分析帧长，以避免规整过程对其产生不必要的平滑，保证规整算法能够适用于不同的信号成分.其次，利用插值的周期，对语音信号进行分段，得到如图 3($b$)所示的分段语音.最后，利用三次 B-样条插值构建规整函数，将各段信号规整为如图 3($c$)所示的具有恒定周期的信号.

由于规整信号的周期大于等于输入信号的最大周期，故信号规整实质上是利用插值运算对不同周期输入信号进行不同尺度的内插.图 4 给出了一段浊音语音规整前后语谱图的对比结果.

图 4 所示结果表明：一方面，规整语音的基频都处于相同的频率，消除了输入信号周期的时变特征，使得相邻周期间的信号冗余完全由信号波形所决定；另一方面，规整信号的时长较输入信号有所增加，但内插操作令规整信号有意义的频域系数都处于低频，使得需量化的频域系数数目与原始信号相近，不会造成额外的比特消耗.

理论上，利用规整函数可以实现对原始信号的完美重建，但插值过程势必会引入计算误差.为了检验这一误差，本文分别对语音和音频信号进行了规整和反规整，并对重建信号进行了信噪比测试.测试结果表明，重建后语音信号的平均信噪比为 60dB，音频信号的平均信噪比为 52dB，由信号规整引入的计算误差不会对后续量化编码产生影响.

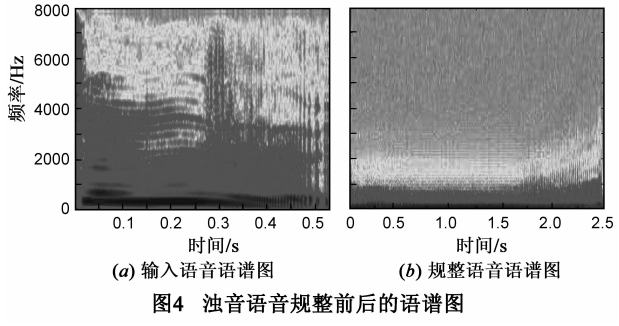



**图4 浊音语音规整前后的语谱图**

## 3 规整信号的稀疏变换

输入信号的规整可以有效去除信号周期的时变性.为进一步去除相邻周期间规整信号波形的相关性，我们将相邻若干周期的规整信号联合组成规整矩阵，进而利用 MLT 和 DCT 在规整矩阵的行和列两个方向分别进行变换，实现对规整信号的稀疏表示.

### 3.1 稀疏变换的基本原理

图 5 给出了构建规整矩阵及稀疏变换的基本原理.规整矩阵 $\boldsymbol{X}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_Q]^{\mathrm{T}}$，其中 $\boldsymbol{A}_k$ 是由规整信号第 $k-1$ 个和第 $k$ 个周期的信号组成的，长度为 $2P$ 的向量，此处 $P$ 为规整信号一个周期波形所包含的样点个数.稀疏变换如下式所示：

$$\begin{aligned} \boldsymbol{H}_{Q\times P} &= \boldsymbol{X}_{Q\times 2P}\boldsymbol{M}_{2P\times P} \\ \boldsymbol{C}_{Q\times P} &= \boldsymbol{D}_{Q\times Q}\boldsymbol{H}_{Q\times P} \end{aligned} \tag{9}$$

式中，$\boldsymbol{M}_{2P\times P}$ 为行变换基矩阵，$\boldsymbol{D}_{Q\times Q}$ 为列变换基矩阵，$\boldsymbol{H}_{Q\times P}$ 为 MLT 系数矩阵，$\boldsymbol{C}_{Q\times P}$ 为 DCT 系数矩阵.因行变换采用 MLT，故行变换基函数[8]为：

$$m(i,j)=w(i)\sqrt{\frac{2}{P}}\cos\Big(\frac{(2i+P+1)(2j+1)\pi}{4P}\Big) \tag{10}$$

式中，$w(i)$ 为窗函数，$i\in[0,2P-1]$，$j\in[0,P-1]$ 分别为行变换基矩阵的行和列索引.因列变换采用 DCT，故列变换基函数[9]为：

$$d(i,j)=c(i)\sqrt{\frac{2}{Q}}\cos\Big(\frac{\pi}{Q}i\Big(j+\frac{1}{2}\Big)\Big) \tag{11}$$

其中，$c(i)=\begin{cases}1/\sqrt{2}, & i=0\\ 1, & i=1,2,\cdots,Q-1\end{cases}$

式中，$i,j\in[0,Q-1]$ 分别为列变换基矩阵的行和列索引.

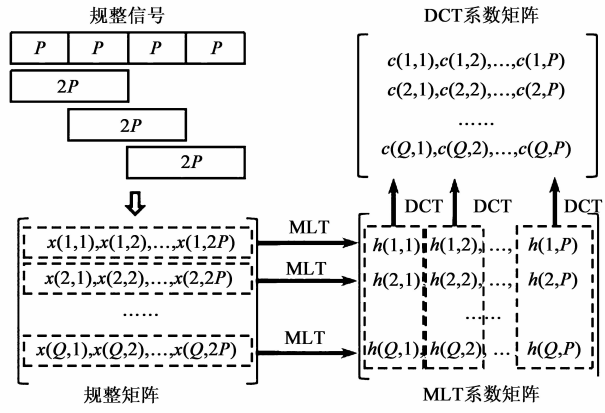


图5　稀疏变换原理图

规整信号相邻周期间波形的相似性,导致 MLT 系数矩阵 $\boldsymbol{H}_{Q\times P}$的行向量,即相邻周期间规整信号的频域系数也极为相似,故通过对 $\boldsymbol{H}_{Q\times P}$在列方向进行 DCT 变换,能够将变换域系数能量集中于少数频点,从而达到数据压缩的目的.

然而,除周期性成分外,规整信号中还包含大量的非周期成份,如图 3($c$)所示的清音语音.若将周期和非周期成分纳入同一个信号矩阵进行变换,一方面会降低对周期成分的压缩效率,另一方面,也会对非周期成分固有的类噪声特性造成损伤,而引入失真.为此,本文提出了一种窗长判决算法,依据不同的信号成分,自适应地选取规整矩阵 $\boldsymbol{X}_{Q\times 2P}$的行向量个数 $Q$,即 DCT 的变换窗长.

### 3.2　自适应窗长判决算法

自适应窗长判决主要基于相邻周期规整信号的相关性,利用各周期信号的能量判决结果对相关系数进行加权,并以加权归一化互相关系数 $\rho_w(k)$作为判决依据,来消除清音段语音对判决结果的不利影响.加权归一化互相关系数 $\rho_w(k)$定义为:

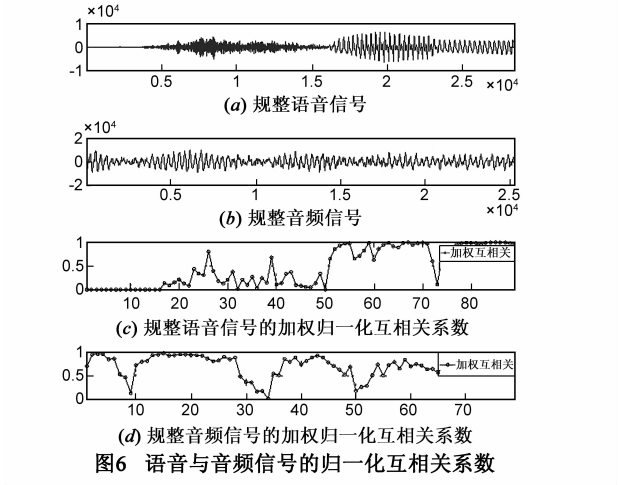
$$\rho_w(k) = Z_k \times \rho(k) \tag{12}$$

式中 $k$ 为周期波形索引,$Z_k$ 为能量判决结果,当信号能量大于一定阈值时为 1,反之为 0,$\rho(k)$为规整信号的第 $k$ 和第 $k+1$ 个周期信号的归一化互相关系数:

$$\rho(k) = \frac{\sum_{n=0}^{P-1} x(n+kP)x(n+(k+1)P)}{\sqrt{\sum_{n=0}^{P-1} x^2(n+kP)x^2(n+(k+1)P)}} \tag{13}$$

图 6 分别给出了规整语音和规整音频的加权归一化互相关系数示例.由规整语音和音频信号计算得到的加权互相关系数在平稳周期成分处具有较高的数值,而在非周期成分处的取值则较小;同时,对于清音和静音段语音相关系数的波动,也具有较好的抑制作用.因此,在加权互相关系数的指导下,我们可实现对分析窗长的有效划分,即假设当前 DCT 的分析窗长为 $Q$(初始值为 1),当第 $k$ 个周期的加权互相关系数大于阈值 $\xi$ 时,则将当前周期信号判定为平稳周期成分,并将其纳入分析窗内,此时调制变换的窗长为 $Q+1$;反之,则将当前周期信号判定为类噪声成分,并开启新的分析窗.

由于非周期成分变化快,因此会被分配以较短的分析窗,以保持信号的非平稳特性,当窗长为 1 时,稀疏变换回归为 MLT 变换;而周期成分变化慢,故被分配以较长的分析窗,以实现变换系数的稀疏表示.



($a$) 规整语音信号
($b$) 规整音频信号
($c$) 规整语音信号的加权归一化互相关系数
($d$) 规整音频信号的加权归一化互相关系数
图6　语音与音频信号的归一化互相关系数

### 3.3　稀疏变换结果分析

图 7 和图 8 分别给出了语音和音频信号的稀疏变换示例.图 7($a$)和图 8($a$)分别为规整后的语音和音频信号,图 7($b$)和图 8($b$)分别为规整信号的 MLT 谱,图 7($c$)和图 8($c$)分别为规整信号经稀疏变换后的 DCT 谱,为直观描述稀疏变换对变换系数能量的集中效果,图 7($d$)和图 8($d$)分别给出了规整信号的 MLT 和 DCT 系数的能量分布.

对于图 7 所示的语音信号而言,稀疏变换通过采用不同的分析窗长,能够将浊音的能量集中于少数周期内,以实现稀疏化,如图 7($d$)所示;而对于清音,则采用了较短的分析窗长,并未产生能量集中的效果,从而避免了对其非平稳特性的不利影响.对于图 8 所示的音频信号而言,其频谱结构相对复杂,相邻周期信号波形变化剧烈,周期成分与非周期成分的区别不如语音信号明显,因此在窗长划分阶段,音频信号大多采用较短的分析窗,以避免对音频中的固有类噪声成分造成过度平滑.

由此可见,稀疏变换通过对信号中的非周期成分进行短窗变换,以保持其固有的类噪声特性,当窗长为 1 时,稀疏变换回归为传统时-频变换方法;而对于周期

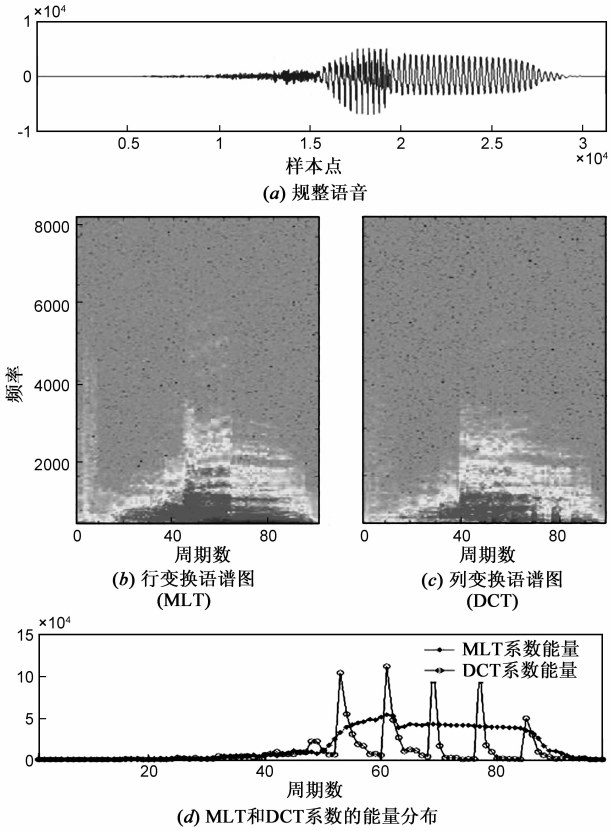



图7 规整语音的稀疏变换系数分布

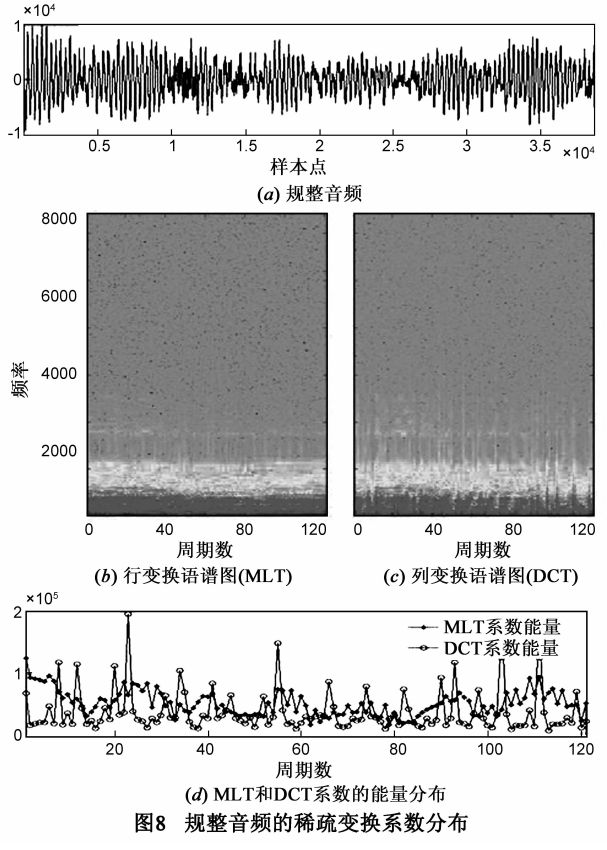



图8 规整音频的稀疏变换系数分布

成分,则利用相邻周期间信号波形的相似性,将稀疏变换后的信号能量集中表现在变换矩阵的少数元素上,进而通过对携带高能量信息的矩阵元素分配充足的量化比特,以提升编码效率.

## 4 编码器框架

图 9 给出了基于信号规整和稀疏变换的语音与音频编码器原理框图.所提编码模型以 20ms 为一帧对输入的宽带信号进行处理,并采用分层编码的方式分别在 24kbps 和 32kbps 码率下实现对语音和音频信号的编码,即在编码核心层对信号的周期、稀疏变换分析窗长、分类控制比特、稀疏变换系数等参数进行标量量化和矢量哈夫曼编码,形成码率为 16kbps 的核心层码流;而在增强层,则对核心层本地解码信号与原始信号残差的 MLT 系数进行标量量化,最终形成码流.

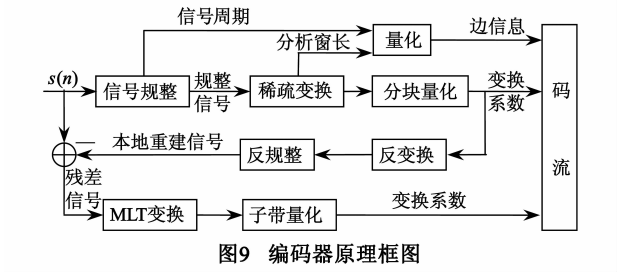



图9 编码器原理框图

### 4.1 编码参数的量化

由于稀疏变换采用可变窗长对规整信号进行分析以实现变换系数的稀疏化,故在编码核心层,首先,对稀疏的变换域系数进行分子带量化,对不同能量的系数分配不同的量化比特;其次,对信号周期和分析窗长等信息进行编码,以便于在解码端重建原始信号;最后,对核心层本地解码信号与原始信号的残差的 MLT 系数进行标量量化,具体比特分配如表 1 所示.

表 1 编码器量化比特分配表(50 帧/s)

| 编码层 | 参数 | 编码速率 | |
|---|---|---|---|
| | | 24 kbps | 32 kbps |
| | | Bit/帧 | Bit/帧 |
| 核心层 | 幅度包络索引变换系数细节信息 | 300 | 300 |
| | 分类控制比特 | 4 | 4 |
| | 分析窗长 | 8 | 8 |
| | 周期信息 | 8 | 8 |
| 增强层 | 残差 MLT 系数 | 160 | 320 |
| 合计 | | 480 | 640 |

在核心层,本文首先对输入的宽带信号进行周期估计,并将连续两帧信号的周期信息组成一个二维矢量进行 8 比特的矢量量化[18].

其次,依据量化后的周期信息对输入信号进行规

整,并构建规整矩阵.进而对规整矩阵进行稀疏变换得到变换域系数.

第三,将所得变换域系数,以 20 个为一组分为若干子带,各子带幅度包络经标量量化和矢量哈夫曼编码形成码流.

第四,根据子带幅度包络量化值与编码可利用比特数,得到 16 种编码模式,并从中选取最优分类模式,选出的分类模式标号用 4 比特分类控制信息表示.

第五,根据选取的最优分类方法,对各子带变换域系数的细节信息进行标量量化和矢量哈夫曼编码[19].

最后,利用 8 比特对稀疏变换的分析窗长进行量化.

由于每帧信号周期的不同,导致子带个数随之变化,故核心层最终形成的码流由变长的幅度包络比特、变长的变换系数细节信息比特、4 分类控制比特、8 分析窗长控制比特和 8 周期信息比特 5 个部分组成.

增强层的量化则相对简单:在完成核心层编码后,编码器在本地重建量化后的信号并与原始信号做差,对残差信号进行 MLT 变换后,对频域系数进行标量量化,以补偿由周期估计、分析窗长划分的误差所导致的失真,以此恢复过渡段音频的暂态特性,提升合成语音和音频的主观听觉质量.

### 4.2　系统解码

与编码算法相对,解码过程也分为 3 个主要部分进行:首先,在调制变换窗长的指导下,对量化后的稀疏变换系数进行反变换,得到量化后的规整信号;其次,利用信号的周期信息,采用与编码端相同的插值方法,得到规整函数,通过规整函数确定的对应关系,反规整得到时域信号;最后,利用残差信号 MLT 系数,进行反变换得到残差信号的时域样本,与反规整得到的时域信号叠加,获得最终的重建语音和音频.

## 5　质量评测

为验证所提算法的有效性,本文分别选取基于自适应变换熵编码技术的 ITU-T G.722.1[20]语音与音频编码标准,和以 CELP 技术为核心的 AMR-WB[21]自适应多码率宽带语音编码器为参考项.分别对 16kHz 采样的宽带语音和音频信号进行码率为 24kbps 和 32kbps 的编码处理.从 MPEG 数据库和 NTT 中文数据库中选取音频信号、语音信号以及语音与音频混合信号各 8 段作为测试语料,对编码器进行客观和主观测试.

本文首先选用 ITU-T 的 PEAQ[22]和 PESQ[23]测试工具进行客观评测,反映 PEAQ 测试结果的主要参数是客观差异等级(Objective Difference Grade,ODG).ODG 反映了解码信号与原始信号之间的差异程度,其得分范围为[-4,0],得分为 0 表示差异最小,ODG 得分越小表示差异越大,当 ODG 得分为 -4 时说明编码失真不可忍受、非常可厌.测试结果如表 2 所示.

**表 2　客观 ODG 分**

| 编码速率(kbps) | 编码方法 | 测试语料 | |
|---|---|---|---|
| | | 音乐信号 | 语音/音乐混合信号 |
| 24 | 所提算法 | -3.518 | -3.117 |
| | G.722.1 | -3.602 | -3.369 |
| 23.85 | AMR-WB | -3.821 | -3.452 |
| 32 | 所提算法 | -3.247 | -2.930 |
| | G.722.1 | -3.480 | -3.150 |

PESQ 是基于听觉模型的一种客观语音评估方法,通过使用听觉感知模型来进行语音质量的评估,结合信号时-频特性,给出一个语音质量感性评价得分,该分值与 MOS 的范围相近,输出分值的范围在 -0.5 ~ 4.5 之间,分值越高质量越好.测试结果如表 3 所示.

**表 3　客观 MOS 分**

| 编码速率(kbps) | 编码方法 | 测试语料 | | |
|---|---|---|---|---|
| | | 女声语音 | 男声语音 | 语音/音乐混合信号 |
| 24 | 所提算法 | 3.940 | 3.856 | 4.053 |
| | G.722.1 | 3.720 | 3.713 | 3.625 |
| 23.85 | AMR-WB | 3.860 | 4.103 | 3.844 |
| 32 | 所提算法 | 4.061 | 3.977 | 4.216 |
| | G.722.1 | 3.763 | 3.816 | 3.764 |

客观测试结果表明,本文所提编码器在 24kbps 和 32kbps 码率下对语音、音频及混合信号的重建效果均优于同码率下的 ITU-T G.722.1 参考编码器;由于 AMR-WB 编码器的最高速率为 23.85kbps,因此,本文算法仅在 24kbps 码率下与之比较.从结果可以看出,本文所提算法,仅在处理男声语音时略低于 AMR-WB 编码器,对于女声语音、音乐及混合信号的处理效果均要优于参考编码器.

最后,本文将所提编码器分别与编码器进行主观 A/B 测试.测试选取了 12 名对测试语料没有先验知识的听力正常的测听者参与完成,结果如表 4 所示.

以上结果表明,本文所提算法在 24kbps 和 32kbps 码率下,对于语音、音频及其混合信号的合成效果均高于 ITU-T G.722.1 编码器,尤其是对于语音与音频混合信号的编码质量,较参考编码器有明显的提升;在 24kbps 码率下,本文算法对于语音信号的处理效果略好于 AMR-WB,而对音频信号和混合信号而言,由于 AMR-WB 编码器以 CELP 技术为核心,导致合成音频存在明显的可感知失真,因此本文算法在该码率下的合成音质要远远优于参考编码器.

表 4 主观 A/B 测试结果

| 参考算法 | 码率(kbps) | 偏爱算法 | 测试语料 | | | | |
|---|---|---|---|---|---|---|---|
| | | | 音频信号 | 男声语音 | 女生语音 | 混合信号 | 总体平均 |
| ITU-T G.722.1 | 32 | 所提算法 | 55.5% | 11.0% | 4.0% | 40.8% | 27.825% |
| | | G.722.1 | 7.5% | 4.0% | 4.0% | 4.0% | 4.875% |
| | | 无偏爱 | 37.0% | 85.0% | 92.0% | 55.2% | 67.3% |
| | 24 | 所提算法 | 55.5% | 11.1% | 14.8% | 70.4% | 37.95% |
| | | G.722.1 | 7.5% | 7.4% | 11.1% | 7.4% | 8.35% |
| | | 无偏爱 | 37.0% | 81.5% | 74.1% | 22.2% | 53.7% |
| AMR-WB | 24 | 所提算法 | 85.2% | 11.1% | 14.8% | 59.3% | 42.6% |
| | 23.85 | AMR-WB | 0.0% | 7.4% | 11.1% | 3.7% | 5.55% |
| | | 无偏爱 | 14.8% | 81.5% | 74.1% | 37.0% | 51.85% |

上述客观和主观测试结果均表明,本文所提算法能够利用统一模型,实现对语音、音频及其混合信号的有效编码.

## 6 结论

本文基于语音和音频信号的固有周期性特征,分别从输入信号的周期和信号波形两个角度去除冗余,以提升变换域编码的压缩效率.首先,所提编码模型利用规整函数提取输入信号周期的时变特征,并将输入信号规整为具有恒定周期的信号,以消除相邻周期输入信号时间上的相关性;其次,对规整信号构建规整矩阵,通过稀疏变换消除规整信号相邻周期间信号波形的冗余,实现稀疏表示;最后,对表征信号能量大小的稀疏矩阵元素进行不同精度量化编码,实现语音和音频信号的通用编码.该模型不同于传统的基于选择机制的通用编码器,无需对输入信号类型进行判别,在根本上消除了由于类型判别和模式切换错误所导致的失真问题.本文通过对宽带语音和音频信号进行初步实验和测试,证明所提编码器对语音和音频信号都能取得较好的编码效果,验证了这一算法的可行性和有效性.

### 参考文献


[1] 鲍长春.数字语音编码原理[M].西安:西安电子科技大学出版社.2007.
Bao Chang-chun. Principles of Digital Speech Coding[M]. Xi'an, China: Xidian University Press. 2007. (in Chinese)

[2] Xiao-ming Li, Chang-chun Bao, W Bastiaan Kleijn. Speech coding based on pitch synchrony and two-stage transformation [A]. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP2013) [C]. Vancouver, Canada: IEEE, 2013. 8159 – 8163.

[3] Takehiro Moriya. Technologies forspeech and audio coding [A]. Proceedings of the IEEE International Symposium on Consumer Electronics [C]. Kyoto, Japan: IEEE, 2009. 148 – 149.

[4] ITU-T G.729.1. An 8-32 kb/s Scalable Wideband Coder Bitstream Interoperable with G.729[S]. 2006 – 05.

[5] 贾懋▮,鲍长春.一种符合 ITU-T 指标要求的嵌入式立体声语音频编码新方法[J].电子学报,2009,37(10):2291 – 2297.
JIA Mao-shen, BAO Chang-chun. An embedded stereo speech and audio coding method meeting the requirements of ITU-T terms of reference[J]. Acta Electronica Sinica, 2009, 37(10): 2291 – 2297. (in Chinese)

[6] ITU-T G.718. Frame Error Robust Narrowband and Wideband Embedded Variable Bit-rate Coding of Speech and Audio from 8 – 32 kb/s[S]. 2008.

[7] 3GPP. TS 26.290 V6.3.0. Extended Adaptive Multi-Rate-Wideband(AMR-WB+)Codec[S]. 2005 – 6.

[8] H Malvar. Lapped transforms for efficient transform/subband coding[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1990, 38(6): 969 – 978.

[9] N Ahmed, T Natarajan, K R Rao. Discretecosine transform[J]. IEEE Transactions on Computers, 1974, C-23(1): 90 – 93.

[10] 刘靖宇,鲍长春,李如玮.基于离散余弦变换的波形内插语音编码算法[J].电子学报,2009,37(7):1599 – 1605.
LIU Jing-Yu, Bao Chang-chun, LI Ru-wei. Waveform inerpolation speech coding based on DCT[J]. Acta Electronica Sinica, 2009, 37(7): 1599 – 1605. (in Chinese)

[11] Ted Painter, Andreas Spanias. Perceptual coding of digital audio[J]. Proceedings of the IEEE, 2000, 88(4): 451 – 513.

[12] K Brandenburg, G Stoll, Y Dehery, et al. ISO-MPEG-1 Audio: A generic standard for coding of high-quality digital audio [J]. AES: Journal of the Audio Engineering Society, 1994, 42(10): 780 – 792.

[13] Neuendorf Max, Multrus Markus, et al. MPEG unified speech and audio coding-the ISO/MPEG standard for high-efficiency audio coding of all content types [A]. Proceedings of the 132nd Audio Engineering Society Convention [C]. USA: AES, 2012. 248 – 269.

[14] M Wolters, K Kjorling, D Homm, H Purnhagen. A closer look into MPEG-4 High Efficiency AAC [A]. Proceedings of the 115th AES Convention [C]. New York, USA: AES, 2003. 5871 – 5886.

[15] M Nilsson, B Resch, M Y Kim, W B Kleijn. A canonical representation of speech [A]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing [C]. USA: IEEE, 2007, vol.4. IV849 – IV852.

[16] Resch M Nilsson, A Ekman, W B Kleijn. Estimation of the instantaneous pitch of speech[J]. IEEE Transactions on Speech

Audio Processing,2007,15(3):813 – 822.

[17] M Unser, A Aldroubi, M Eden. B-spline signal processing: Part Ⅰ-theory[J]. IEEE Transactions on Signal Processing, 1993,41(2):821 – 833.

[18] M Unser, A Aldroubi, M Eden. B-spline signal processing: Part Ⅱ-efficient design and applications[J]. IEEE Transactions on Signal Processing,1993,41(2):834 – 848.

[19] 薛二娟,鲍长春,李如玮. 基于二维非负矩阵分解的1kb/s WI语音编码算法[J]. 电子学报,2010,38(7):1574 – 1579.
XUE Er-juan, BAO Chang-chun, LI Ru-wei. 1kb/s waveform interpolative speech coding based on two-dimensional nonnegative matrix factorization[J]. Acta Electronica Sinica,2010,38(7):1574 – 1579. (in Chinese)

[20] ITU-T Recommendation G.722.1 Low-Complexity Coding at 24 and 32 kbit/s for Hands-Free Operation in Systems with Low Frame Loss[S]. Geneva,2005 – 05.

[21] 3GPP TS 26.171. Adaptive Multi-Rate-Wideband(AMR-WB) Speech Codec;General Description[S]. 2002.

[22] F Baumgarte, A Lerch. Document 6QI18-E. Implementation of Recommendation ITU-R BS. 1387, Delayed Contribution [S]. February 2001.

[23] ITU-T Recommendation P. 862. Perceptual Evaluation of Speech Quality(PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codes[S]. 2001.

## 作者简介

**李晓明**　男,1983 年生,内蒙古赤峰人,北京工业大学博士研究生,主要研究方向为语音与音频编码.
E-mail:lixiaoming@emails.bjut.edu.cn

**鲍长春**　男,1965 年生,内蒙古赤峰人,博士,北京工业大学教授、博士生导师,IEEE 高级会员,国际语音通信学会(ISCA)会员,亚太信号与信息处理学会(APSIPA)会员,中国电子学会理事,中国声学学会理事,信号处理分会委员,《信号处理》和《数据采集与处理》编委. 主要研究方向为语音与音频信号处理.
E-mail:baochch@bjut.edu.cn