基于匹配跟踪的感知梯度正弦建模方法*

张文耀+, 许 刚, 王裕国

(中国科学院 软件研究所,北京 100080)

A Sinusoidal Modeling Method Based on Matching-Pursuits with Perceptual Gradient

ZHANG Wen-Yao+, XU Gang, WANG Yu-Guo

(Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: 86-10-82660088 ext 693, E-mail: zhwenyao@263.net; zhangwy@sinosoftgroup.com

http://www.iscas.ac.cn

Received 2001-11-15; Accepted 2002-02-26

Zhang WY, Xu G, Wang YG. A sinusoidal modeling method based on matching-pursuits with perceptual gradient. *Journal of Software*, 2003,14(3):467~472.

Abstract: As an adaptive algorithm of signal decomposition, matching pursuits provides a new framework for sinusoidal modeling of speech and audio signal. In this paper, the procedure of sinusoidal modeling using matching pursuits is analyzed as well as the sinusoidal modeling algorithm using perceptually weighted matching pursuits. And a method of sinusoidal modeling with perceptual gradient is proposed. The proposed method, which adopts the adaptive feature of matching pursuits, computes dynamically a masking threshold from the currently synthesized signal using the psychoacoustic model. With the threshold, it extracts the most perceptually significant component from the residual signal. Therefore, the perceptual information contained in the synthesized signal increases as quickly as possible. The quality of the synthesized speech by this approach is rather high even if the model precision is low. Experiments prove that the method in this paper uses the features of hearing system in a better way, and the modeling is reasonable and efficient. Both the objective compare of SNR and the subjective listening test show the rationality and superiority of the new method.

Key words: sinusoidal modeling; matching pursuit; perceptual gradient; psychoacoustic model; speech signal processing

摘 要: 匹配跟踪作为一种自适应的信号分解算法,为语音和音频正弦建模提供了一个新的框架.分析了基于匹配跟踪的正弦建模过程以及感知加权匹配跟踪正弦建模算法,并在此基础上提出了感知梯度正弦建模方法.该方法结合匹配跟踪自适应的动态特征,利用心理声学模型计算当前合成信号的动态掩蔽阈值,以此为参考提取残差信号中感觉最明显的信号分量,从而最大限度地增加合成信号中的感知信息.在模型精度不高的情况下,该方法也能得到合成质量比较高的语音.实验表明,该方法更好地利用了人耳的听觉特性,建模结果更为合理、有效.客观的信噪比和主观试听测试都显示了所提出算法的合理性与优越性.

关键词: 正弦建模;匹配跟踪;感知梯度;心理声学模型;语音信号处理

---

* **第一作者简介**: 张文耀(1974－),男,江西萍乡人,博士生,主要研究领域为语音信号处理,模式识别.

中图法分类号: TP391 文献标识码: A

正弦建模(sinusoidal modeling)在语音和音频信号处理中得到了广泛的应用,如语音压缩[1,2]、语音变换[3]、语音合成以及说话人分离[4]等.其目标是为输入信号 $s(n)$ 寻找 K 个正弦信号分量,由这 K 个正弦分量之和构成对原信号的一种逼近,即

$$s(n) \approx \hat{s}(n) = \sum_{k=1}^{K} a_k \cos(\varpi_k n + \phi_k),\ n = 0,1,...,N-1, \tag{1}$$

其中 a_k, ϖ_k 和 ϕ_k 分别表示正弦分量的幅度、频率和相位参数.如果不对式(1)的正弦参数施加约束的话,不同的实现技术会得到不同的正弦建模结果.

传统的正弦建模方法是在短时 Fourier 变换(short-time Fourier transform,简称 STFT)的基础上,采用谱峰值检测技术得到模型参数[1,2].另外就是采用基于帧的分析合成技术[5].尽管两种方法各有优势,但是分析合成技术采用最小化均方误差准则,每步迭代都从输入信号中查找并删除能量最大的正弦分量(sinusoidal component),保证了模型的收敛性.但是,对于人耳听觉感知来讲,能量最大的信号分量不一定是感觉最明显的.由此,文献[6]提出了一种新的分析合成正弦建模方法.该方法采用基于帧的感知加权匹配跟踪技术,结合人类听觉系统的心理声学特征,每步迭代都查找并删除输入信号中感觉最明显的正弦信号分量.为此,首先根据心理声学模型(psychoacoustic model)[7,8]计算输入信号的全局掩蔽阈值(global masking threshold),以此作为匹配跟踪字典的加权序列,然后运用匹配跟踪(matching pursuit,简称 MP)算法[9]提取各个正弦信号分量的参数.实际上,就是从当前残差信号中提取信号掩蔽率(signal-to-mask ratio,简称 SMR)最大的信号分量.其中掩蔽阈值为初始输入信号的全局掩蔽阈值,在分解迭代过程中保持不变.这种固定掩蔽阈值的 SMR 选择方法与匹配跟踪动态自适应的分解思想并不十分吻合,结果在正弦信号分量不是足够多的情况下,合成语音失真严重,清晰度和可懂度都不高.

为此,本文提出基于匹配跟踪的感知梯度正弦建模算法.该方法将输入信号分解成两部分:当前的合成信号与当前的残差信号.每步分解迭代过程都从当前残差信号中选择相对于当前合成信号感觉最明显的正弦分量.虽然其分量选择准则也是基于 SMR,但其中的掩蔽阈值是从当前合成信号的计算中得到的,在跟踪过程中是动态变化的,与匹配跟踪的自适应特征一致.实验结果表明,本文方法取得了明显的改善效果,在同等条件下不仅合成的语音质量优于文献[6],而且客观的信噪比也有较大的提高.

本文首先简要介绍基本的匹配跟踪思想,然后分析基于匹配跟踪的正弦建模过程以及感知加权匹配跟踪正弦建模算法,并在此基础上给出本文的感知梯度正弦建模过程.接着,为检验本文方法的性能进行了对比实验,并详细分析了实验结果.最后是本文的结论以及进一步的研究方向.

1 基本的匹配跟踪算法

匹配跟踪是一种自适应的信号分解迭代算法.它在一个高度冗余的字典(dictionary)空间 D 中将输入信号 $s(n)$ 分解成一组原子(atom)信号的线性组合[9].假定包含 M 个原子的字典为

$$D = \{g_m\};\ m = 0,1,...,M-1. \tag{2}$$

并且每个原子都具有单位范数,即 $\| g_m \| = 1$.匹配跟踪的分解迭代过程如下:

设置初始输入信号为当前残差信号,即令 $r_0 = s(n)$.在第 $k\ (k \geq 0)$ 步迭代中,查找第 k 个原子索引 m_k,使该原子与当前残差信号 r_k 的相关系数最大.此时,通过正交投影可以得到:

$$r_k = \left\langle r_k, g_{m_k} \right\rangle g_{m_k} + r_{k+1}, \tag{3}$$

其中 $\left\langle r_k, g_{m_k} \right\rangle$ 表示两个向量的内积, r_{k+1} 为新的残差信号,

$$m_k = \arg\max_m \left(\left| \left\langle g_m, r_k \right\rangle \right| \right), \forall m. \tag{4}$$

由于采用了正交投影,所以

$$| r_k |^2 = \left| \left\langle r_k, g_{m_k} \right\rangle \right|^2 + | r_{k+1} |^2 . \tag{5}$$

继续这种分解直到第 K 步,并令 $\alpha_k = \langle r_k, g_{m_k} \rangle$,将得到

$$s = \sum_{k=0}^{K-1} \alpha_k g_{m_k} + r_K. \tag{6}$$

从式(3)~式(6)可见,匹配跟踪每次都从残差信号选取能量最大的信号分量.随着分解次数的增加,式(6)右端原子向量的线性组合可以任意地逼近原始信号.但是匹配跟踪过程通常在满足某种精度条件时就终止了,如残差能量低于某一阈值,具体情况需根据应用场合而定.

2 基于匹配跟踪的正弦建模

选用不同的原子字典,可以将匹配跟踪应用到不同的场合[9~11].如果考虑如下复指数原子组成的正弦字典(sinusoidal dictionary):

$$D_S = \{g_m(n) = c_m \mathrm{e}^{j\varpi_m n}; m = 0,1,...,M-1\}, \tag{7}$$

其中 c_m 为原子的归一化系数,ϖ_m 为正弦原子的频率参数.此时,匹配跟踪类似于基于帧的分析合成正弦建模.因为分解迭代的结果是一组正弦波形的加权和.此处字典空间是由复指数原子所刻画的,而实际中面临的通常是实信号.为了处理方便,可以采用共轭子空间投影技术[12],在由字典原子及其复共轭所形成的子空间中计算相关系数,其结果也以共轭对的形式出现,这样第 k 步迭代得到的残差信号为

$$r_{k+1}(n) = r_k(n) - \alpha_k g_{m_k} - \alpha_k^* g_{m_k}^* = r_k(n) - a_k \cos(\varpi_k n + \phi_k). \tag{8}$$

合成信号(或逼近信号)则是

$$\hat{s}_K = \sum_{k=0}^{K-1} a_k \cos(\varpi_k + \phi_k), \tag{9}$$

与式(1)中的正弦模型统一起来了.

如前所述,匹配跟踪选择的是当前残差信号中能量最大的信号分量.而在语音和音频信号处理中,由于人耳听觉感知的非线性特征,能量最大的信号分量不一定是感觉最明显的.为此,文献[6]将心理声学模型引入匹配跟踪过程,提出基于感知加权的匹配跟踪正弦建模算法.该方法首先根据心理声学模型计算初始输入信号的全局掩蔽阈值,以此作为正弦原子字典的加权序列,然后将普通内积修改为加权内积,再按照匹配跟踪的框架进行正弦建模.其实质是选取当前残差信号中感知能量最大的分量(即感觉最明显的)添加到当前合成信号中.听觉感知程度以信号掩蔽率(signal-to-mask ratio,简称 SMR)来衡量,SMR 越大,感觉越明显.基于 SMR 的分量选择准则较好地利用了人耳的听觉特性.在提取了足够多的信号分量的情况下,虽然合成信号和原始信号之间误差很大,但是两者在听觉感知上几乎等同[6].

感知加权匹配跟踪对于残差信号分量的感知判别是建立在初始输入信号的全局掩蔽阈值的基础上.掩蔽阈值在整个匹配跟踪过程中是不变的,但残差信号却不断变化.这相当于每次都从残差信号中选取相对于原始输入信号而言最明显的分量,而不是真正的残差信号中感觉最明显的.同时,没有考虑合成信号的状态和影响.从心理声学掩蔽模型可知,各个原子信号之间存在掩蔽效应,前面选择的原子对后来添加的原子信号将产生掩蔽作用.如果采用全局掩蔽阈值作为原子选择准则,就可能出现后来添加的原子信号被已有的原子信号全部或部分掩蔽的现象,从而降低建模效率.

为此,本文提出基于动态掩蔽阈值的原子选择准则,将掩蔽阈值与当前合成信号关联起来,每次都选择残差信号中相对于当前合成信号感觉最明显的信号分量.这样,每次都最大限度地增加了合成信号的感知信息.由于考虑了当前原子集合的掩蔽效应,所添加的原子相对于现有的原子集合而言感觉是最明显的.从感知上讲,建模效率是最高的.这类似于数学上的梯度概念,借用此概念,称本文的基于动态掩蔽阈值的正弦建模算法为感知梯度正弦建模(sinusoidal modeling with perceptual gradient).

3 基于匹配跟踪的感知梯度正弦建模

由前面的论述可知,选择正弦字典,采用共轭子空间投影技术,经过 K 步分解,输入信号 s 可以表示成:

$$s(n)=\hat{s}_K(n)+r_K(n), \tag{10}$$

其中 $\hat{s}_K$ 为所选原子集合构成的合成信号,如式(9)所示; r_K 为残差信号.正弦建模的关键在于估计 $\hat{s}_K$ 中正弦信号分量的参数.令第 k 个分量的参数集合为 $A_k=(a_k,\omega_k,\phi_k)$,感知梯度建模过程为

(1) 根据输入信号的时频分辨率建立式(7)所示的正弦原子字典.

(2) 令 $r_0=s$,并计算初始掩蔽阈值 T_0 .由于初始时刻 $\hat{s}_0$ 为空,因此设置 T_0 为安静时刻的绝对掩蔽阈值[8,9].

(3) 进入第 $k(k\geq 0)$ 步分解迭代过程,查找残差信号 r_k 中感觉最明显的正弦分量.为此,首先对 r_k 进行 DFT 变换,得到其能谱密度 P_k ,再根据 SMR 最大化准则确定原子 A_k 的频率成分 ω_k ,即

$$\omega_k=\arg\max_{\omega}\left(\frac{P_k(\omega)}{T_k(\omega)}\right). \tag{11}$$

(4) 利用匹配跟踪子空间投影技术计算原子 A_k 的幅度 a_k 和相位 ϕ_k ,并更新合成信号和残差信号,即

$$\hat{s}_{k+1}(n)=\hat{s}_k(n)+a_k\cos(\omega_k n+\phi_k), \tag{12}$$

$$r_{k+1}(n)=r_k(n)-a_k\cos(\omega_k n+\phi_k). \tag{13}$$

(5) 计算新的掩蔽阈值 T_{k+1} .将原子 A_k 当作音调掩蔽音(tonal masker),计算其单独的掩蔽阈值 T_{A_k} ,并迭加到 T_k ,得到新的合成信号 $\hat{s}_{k+1}$ 的掩蔽阈值 T_{k+1} .

(6) 令 $k=k+1$,重复步骤(3)~步骤(5),直到满足终止条件为止.

迭代终止条件可以根据需要设定.最简单的是在达到设定的模型精度,即 $k=K$ 时结束.另一个顺理成章的终止条件是在合成信号完全掩蔽残差信号时终止.因为此时继续分解已经不能改变合成信号的任何感知效果,后续分解已经没有意义.判断是否完全掩蔽只需查看步骤 3 中原子的 SMR 值是否大于 1.如果小于 1,则表明 r_k 已经完全被 $\hat{s}_k$ 掩蔽.虽然均方意义上的残差可能仍然很大,但是重构信号与原始信号在感觉上是等同的.

4 实验结果的分析与对比

为检验本文算法,选择不同采样率、不同发音的语音信号进行了对比实验.参与对比实验的有:基本的匹配跟踪正弦建模算法(basic matching pursuit,简称 BMP)、文献[6]的感知加权匹配跟踪算法(perceptually weighted matching pursuit,简称 PWMP)以及本文的感知梯度正弦建模算法(sinusoidal modeling with perceptual gradient,简称 SMPG).实验中使用了 MPEG-1 的心理声学模型 I[8,9],采用了重叠相加分段处理技术.

实验发现:当选择同样数目的正弦分量且分量数目比较少时,BMP 产生类似低通滤波的效果,低沉、压抑,与原声相比失去了许多高频音色信息.PWMP 则出现音调失真(tonality artifacts),合成的声音好像与另外一个被调制的杂音混合在一起,有时严重影响清晰度.SMPG 避免了这些现象,虽然音质与原音相比也存在距离,但是整个听觉效果比较和谐,音色更宽宏,清晰度也很高.当模型精度不高时,三者的差别尤其明显.但是,当模型原子数目足够多时,三者的感觉差异非常细微,几乎都等同于原始信号.

分析建模过程,发现这些差异的原因在于:BMP 采取能量最大化准则,最初选择的原子分量主要集中在低频区域.这与低频信号的能量通常比较大的现象相吻合.PWMP 采用输入信号的全局掩蔽阈值作为字典原子的加权值,虽然每次都选择感知能量最大的分量,但是由于固定的掩蔽阈值与动态的分解过程不一致,各个原子分量之间的关系不清晰,既不能按能量来组织,也不能按真正的感觉差异来组织,因此建模效果不稳定.SMPG 每次都最大化合成信号的感知信息,依次选择的原子从感知上讲是个有机序列,随着原子数目的增加,语音信息以最快的速度增加.因此在原子数目比较少的情况下也能得到相当好的合成语音.当模型中原子数目足够多时,三者的感觉效果差别不大的原因在于匹配跟踪算法是收敛的.合成信号的能量和频率成分都可以无限逼近原始信号,因此音质也就无限接近.只是三者的收敛方式和收敛速度不同,BMP 是能量收敛最快的,SMPG 是感知收敛最快的.正是这种收敛差异导致了它们的感知差异.

图 1~图 3 分别给出了 3 种算法针对同一段语音信号所选择的前 30 个原子的频率分布.这些原子是按照频率索引以'+'标示在输入信号的能谱图上,旁边的数字表示被选择的先后顺序.图 2 中的虚线是该段输入信号的全局掩蔽阈值.图 3 的虚线是跟踪过程中不同时刻的掩蔽阈值.这簇掩蔽曲线的上边和下边分别表示最终和最初的掩蔽阈值.从图中可以看到,BMP 的原子集中在低频的高能区域;PWMP 受全局掩蔽阈值的影响,原子偏重

某些频率区间;SMPG 的原子分布则比较有规律,基本覆盖整个频率空间,既强调了某些重要的频率区域,也兼顾了高频高能分量.这从频域可以反映它们的感知差异.

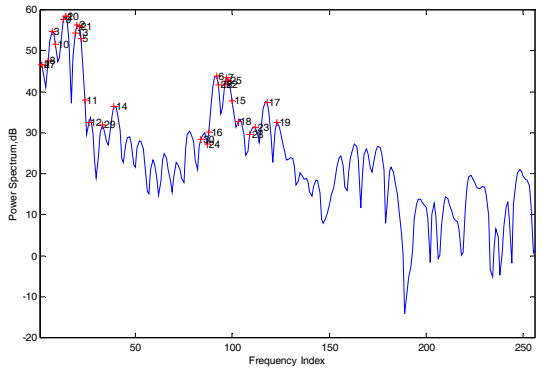

Fig.1 Example of the BMP result
图 1 BMP 结果示例

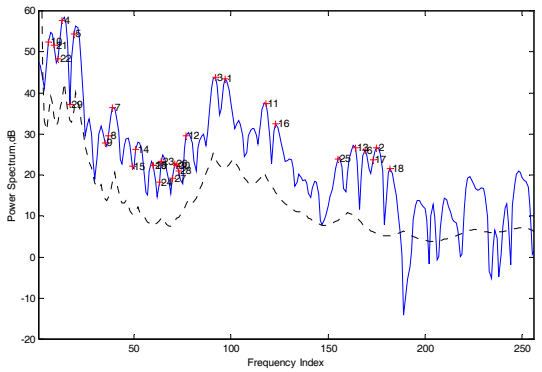

Fig.2 Example of the PWMP result
图 2 PWMP 结果示例

另外,对比建模前后的波形发现,BMP 的信噪比(signal-to-noise ratio,简称 SNR)最高,SMPG 又比 PWMP 高出许多.SMPG 的 SNR 值高于 PWMP,说明 SMPG 能更好地逼近原始信号.这从另一个角度反映了感知梯度正弦建模的优越性.BMP 的高信噪比与其能量最大化准则是一致的.表 1 给出了一段男声话音(“祝你好运”)和女声话音(“Nice to meet you”)在不同模型精度下的 SNR 对比结果.此处,SNR 仍可比较是因为匹配跟踪可以看做是时域的波形逼近.这与传统的谱峰值建模技术不同.传统谱峰值建模的 SNR 值非常低(约 1~2dB)[4],已经没有实际意义.

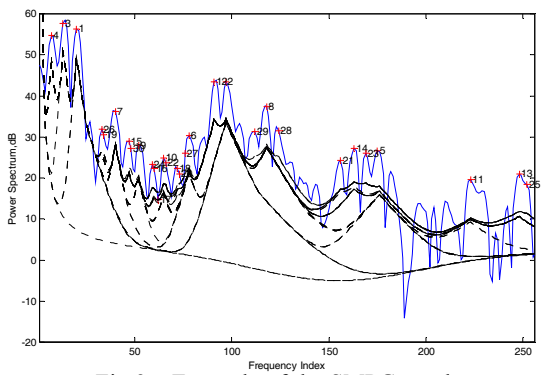

Fig.3 Example of the SMPG result
图 3 SMPG 结果示例

为了进一步检验合成效果,选取了 12 句日常生活中的中英文短语,邀请了 10 位试听者(男女各半)参加 A/B 对比测听实验.结果是,所有试听者都认为 SMPG 优于 PWMP 和 BMP,而认为 PWMP 优于 BMP 的只占 74.6%.

本文的感知梯度算法则结合了时域和频域的特征:匹配跟踪可以看做是时域的波形逼近,感知心理声学模型则是从频域来考虑的.感知梯度与匹配跟踪的有机结合达到了比较满意的建模结果.大量的实验表明,本文的算法是合理的、有效的.

Table 1 SNR comparison among the three methods under different models' precision
表 1 3 种方法在不同模型精度下的 SNR 比较

		K=10	*K*=20	*K*=30
Male voice	BMP	14.934 7	19.832 6	23.055 1
	PWMP	6.354 3	7.662 0	8.688 2
	SMPG	8.505 7	12.039 3	13.698 1
Female voice	BMP	17.112 3	21.908 1	25.001 2
	PWMP	6.729 8	8.863 8	9.402 6
	SMPG	13.996 8	16.276 6	16.769 9

5 结 论

匹配跟踪作为一种自适应的信号分解算法,为语音和音频信号正弦建模提供了一个新的框架.本文分析了基于匹配跟踪的正弦建模过程以及感知加权匹配跟踪正弦建模算法.基本的匹配跟踪正弦建模建立在能量最大化的基础上,没有考虑心理声学特征.感知加权匹配跟踪虽然考虑了听觉感知特征,但其固定的掩蔽阈值与匹配跟踪的动态过程不太吻合,建模结果不理想.本文提出的感知梯度正弦建模算法结合匹配跟踪自适应的动态

特征,利用心理声学模型,计算出当前合成信号的声学掩蔽阈值,以此为参考提取当前残差信号中感觉最明显的信号分量,每次都以最大的可能增加合成信号的感知信息,从而在模型精度不高的情况下也能得到合成质量比较高的语音.客观的 SNR 比较和主观试听测试都表明了本算法的合理性与优越性.

感知梯度正弦建模较好地利用了心理声学特征,提高了建模效率,其所选择的原子在感知意义上成为一个有机的关联序列.这在实际应用中非常有利,比如在语音或音频压缩中就可以利用其建模结果实现变位率编码,高位率选用较多的原子,低位率使用较少的原子,使每种位率都包含尽可能多的语音感知信息.这种类似于嵌入式编码的灵活机制,对于互联网上音频流的传输以及网络服务质量的提高是相当有利的.其具体工作还有待于深入、细致的研究.另外,本文使用的心理声学模型比较简单,为了更好地利用听觉掩蔽等声学现象,需要研究更为准确、更为复杂的心理声学模型.

References:

[1] McAulay RJ, Quatieri TF. Speech analysis-synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech and Signal Processing, 1986,34(4):744~754.

[2] McAulay RJ, Quatieri TF. Sinusoidal coding. In: Kleijin WB, ed. Speech Coding and Synthesis. Netherlands: Elsevier Science B.V., 1995. 123~173.

[3] George EB, Smith MJT. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. IEEE Transactions on Acoustics, Speech and Signal Processing, 1997,5(5):389~406.

[4] Morgan DP, George EB, Lee LT, Kay SM. Co-Channel speaker separation by harmonic enhancement and suppression. IEEE Transactions on Acoustics, Speech and Signal Processing, 1997,5(5):407~425.

[5] George EB, Smith MJT. Analysis-by-Bynthesis/Overlap-Add sinusoidal modeling applied to the analysis and synthesis of musical tones. Journal of the Audio Engineering Society, 1992,40(6):497~515.

[6] Verma TS, Meng HY. Sinusoidal modeling using frame-based perceptually weighted matching pursuits. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Piccataway, N.J.: IEEE, 1999. 981~984.

[7] ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3. Information Technology——Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbits/s, Part3: Audio. 1992.

[8] Painter T, Spanias A. Perceptual coding of digital audio signal. Proceedings of the IEEE, 2000,88(4):451~513.

[9] Mallat S, Zhang Z. Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing, 1993,41(12): 3397~3415.

[10] Jaggi S, Karl WC, Mallat S, Willsky AS. High resolution pursuit for feature extraction. Applied and Computational Harmonic Analysis, 1998,5(3):428~449.

[11] Zakhor NR. A very low bit-rate video coding based on matching pursuits. IEEE Transactions on Circuits and Systems for Video Technology, 1997,7(1):158~171.

[12] Goodwin M, Vetterli M. Matching pursuit and atomic signal models based on recursive filter banks. IEEE Transactions on Signal Processing, 1999,47(7):1890~1902.