2001年度 第14回 信號處理合同學術大會 論文集 第14卷 1號

음성 및 음악을 위한 저 전송률 다중모드 하모닉 변환 여기 부호화기

김종학, 이인성
충북대학교 전파공학과

Low Bit Rate Multi Mode Harmonic Transform Excitation Coding for Speech and Music

Jonghark Kim, Insung Lee
Dept, of Radio Engineering, Chungbuk National Univ.

요 약

본 논문은 음성 및 음악을 위한 새로운 4kbps 다중모드 하모닉 변환 여기 부호화 방법을 제안한다. 제안된 부호화방법은 음성/음악 분류기에 의해 분류된 신호를 각각 하모닉-잡음 여기모델과 MLT 여기모델로 부호화한다. 하모닉-잡음 여기모델에서는 전이구간과 유/무성음 혼합신호의 모델링오차 개선을위해 MP(Matching Pursuit)방법과 혼합된 잡음스펙트럴을 표현하기위한 캡스트럼 LPC 잡음 모델, 빠른 정현파 합성법을 제안한다. 음악에서는 비트할당 효율을 높이기위한 LP 적응 피크 분석을 적용한 MLT(Modulated Lapped Transform) 부호화 방법을 제안한다. 제안된 방법을 적용한 4kbps 음성부호화 방법은 전이구간에서의 향상된 모델링 구조를 보여주었으며, 주관적음질 평가 8kbps QCELP 보다 MOS 0.2 정도 향상된 결과를 얻었다.

1. 서 론

최근 무선 멀티미디어, 패킷전송을 위한 인터넷 응용 및 오디오/비디오 통신회의 같은 몇가지 응용에 대해서 음성과 오디오에 대한 관심이 증가되고 있다. 이러한 응용 서비스들은 음성 및 음악이 혼합된 다양한 신호에 대해서도 음질을 보장하도록 요구하고 있으며, 그 연구 부분에서도 음성 및 오디오에 대해 모델화된 압축 알고리즘들이 개발되고 있다.

많은 음성 및 오디오 부호화기에 대해서, 낮은 전송률에서의 부호화 알고리즘들은 제한된 입력 신호에 대해서만 디자인 되어있기 때문에 다양한 신호에 대한 기대 음질을 만족하지 못한다. 최근 까지, 양 음성, 음악 신호에 대해 디자인된 알고리즘들은 그리 주목할만한 주의를 끌지 못하였다[1][2].

특히, 하모닉 부호화기는 CELP 부호화기와는 달리 주파수 영역 검색을 적용한 간단한 추출 구조을 사용하기 때문에 복잡도 및 음질 측면에서 좋은 성능을 보여주었으나, 단지 기본 주파수에 의존하는 하모닉 구조는 음악에서는 만족할만한 음질을 얻기 힘들다. 비록 피크 연속 과정과 같은 방법이 첨가가 된다 할지라도, 단지 하모닉 성분들의 합에의해 모델화된 여기신호는 오디오신호를 표현하는데에 큰 제한을 가진다.

본 논문에서는 이러한 양 음성, 음악에 대하여 만족할만한 음질을 음질을 얻을 수 있는 혼합구조에 대하여 소개한다. 제안된 알고리즘은 하모닉 잡음 부호화기에서 MP(Matcing Pursuit) 스펙트럴 크기 분석방법 및 캡스트럼 LPC 잡음 분석방법과 빠른 정현파 합성방법을 포함하며 음악에서는 LP 적응 피크 분석을 적용한 MLT(Modulated Lapped Transform) 부호화 방법을 포함한다. 제안된 부호화기의 기본 구조는 2장에서 설명되며, 음성/음악 분류기는 3장에서 설명된다. 하모닉 잡음 여기 부호화기와 MLT 여기 부호하기는 4, 5장에서 설명된다. 비트할당을 포함한 실험결과에 대한 설명을 6장에서 전개한다.

2. MMHTC의 전체 구조

제안된 부호화기와 복호화기의 블록다이어그램이 그림 1과 2에 나타나있다. 그림 1에 보여지는 것처럼, 그 알고리즘은 하모닉 여기 모드 및 피치 분석과정이 없는 CELP, 변환 부호화 여기 모드인 3가지 부호화기 모드로 구성된다. 각 모드는 음성/음악 분류기 및 유/무성음 분류기가 각모드를 결정하기 위해 사용된다. 먼저 20msec 당 LP(Linear Prediction) 분석이 적용되며 그 잔여신호가 각 모드에 따른 여기모델에 의해 부호화 된다. 음성에서의 유성음은 정현파 모델에의한 하모닉 부호화기에 의해 부호화 된다. 하모닉 부호화기의 중요특징 블록은 MP(Matching Pusuit) 분석 부분, 하모닉 양자화기, IFFT(Inverse Fouriour Transform)를 사용한 빠른 합성부분으로 구성된다. 또한, 하모닉 스펙트럴에 혼합된 무성음 성분의 신호를 효과적으로 표현하기 위해 캡스트럼 LPC 잡음 스펙트럴 분석기를 사용한다. 복호화기에서의 하모닉 합성법은 그 복잡도를 줄이기 위해 IFFT를 사용한 합성법을 사용한다. 음악 신호에서의 변환 부호화기는 20msec 마다 MLT를 적용하며, 저 전송률에서 부족한 비트수를 고려한 LP 적응 피크 검출법을 사용한다.

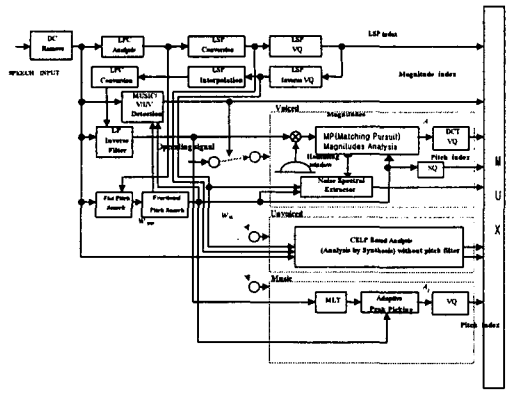


그림 1. MMHTC 부호화기 블록다이어그램

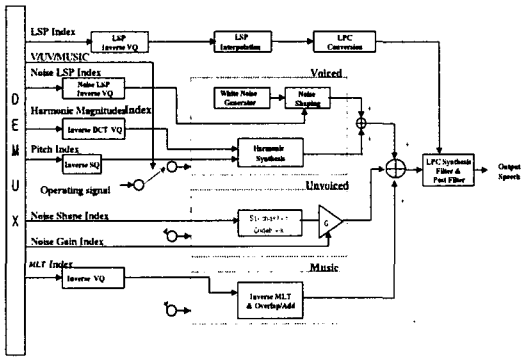


그림 2. MMHTC 복호화기 블록다이어그램

3. 음성/음악 분류기

음성 파형은 큰 크기값의 의사주기 유성음 부분의 매우 규칙적인 패턴을 가진다. 또한 그 스펙트럴 라인에서의 에너지 값은 시간에 대해서 매우 큰 분산 값을 갖는다. 다시말해, 음악신호에서의 하모닉 성분은 음성에서보다 정체적인 특성을 갖는다. 이러한 특성은 음성 및 음악을 구별하기 위한 특징파라미터로 쓰여질 수 있다. 본 제안된 부호화기에서의 음성/음악 분류기는 MPEG4 방식에 바탕을 두며[3], 에너지 분산값과 입력신호의 자기상관값에 대한 첫 번째와 두 번째 피크비율에 대한 분산값의 범위를 체크하여 판별한다..

4. 음성을 위한 하모닉 잡음 부호화기

4.1 MP(Matching Pursuit)을 사용한 하모닉 스펙트럴 크기값 추출법

MP는 과잉완전(Overcomplete) 정현파 사전에서의 밀집(Compact)형 해인 하모닉 스펙트럴 크기값을 구하기 위한 부 최적(Sub-optimal) 순환(Iterative) 알고리즘이다. $g_i[n]=d_{m(i)}[n]$의 과잉완전 사전을 정의하면, $i+1$번째의 잔여신호는 i번째의 잔여신호에서 i번째 선택된 최적 사전요소 성분값 $\alpha_i d_{m(i)}[n]$을 뺀 값이 된다. 여기서 0번째 신호는 원본 신호 값이 되며, i가 무한대로 간다면 사전요소를 이용한 복원신호는 동일한 원본신호를 합성할 수 있다[4].

$$r_{i+1}[n]=r_i[n]-\alpha_i d_{m(i)}[n] \tag{1}$$

최적 파라미터 값을 구하는 과정은 $i+1$번째 신호의 놈(Norm)을 최소화 되도록 하는 해를 구하는 과정이다. 벡터 $\mathbf{r}_{i+1}$와 $\mathbf{g}_i$의 직교(Othogonal) 성질을 이용하면 식 (4)와 같이 α_i를 최대화하는 $\mathbf{r}_i$값을 찾는 과정으로 요약된다[4].

$$\mathbf{g}_i=\arg\min_{\mathbf{g}_i\in\mathbf{D}}\|\mathbf{r}_{i+1}\|^2=\arg\min_{\mathbf{g}_i\in\mathbf{D}}\|\mathbf{r}_i-\alpha_i\mathbf{g}_i\|^2 \tag{2}$$

$$\alpha_i=\frac{\langle\mathbf{g}_i,\mathbf{r}_i\rangle}{\langle\mathbf{g}_i,\mathbf{g}_i\rangle}=\frac{\langle\mathbf{g}_i,\mathbf{r}_i\rangle}{\|\mathbf{g}_i\|^2}=\langle\mathbf{g}_i,\mathbf{r}_i\rangle \tag{3}$$

$$\|\mathbf{r}_{i+1}\|^2=\|\mathbf{r}_i\|^2-|\alpha_i|^2 \tag{4}$$

여기서, $\langle\rangle$은 벡터 적(Inner product)을 뜻한다. 이러한 과정은 큰 복잡도를 요구하지만, α_i 또한 식 (5)와 같이 순환식으로 구현될수 있으며, $\langle\mathbf{g},\mathbf{g}_i\rangle$는 미리 계산될수 있어 그 복잡도는 크게 감소될 수 있다[4].

$$\langle\mathbf{g},\mathbf{r}_{i+1}\rangle=\langle\mathbf{g},\mathbf{r}_i\rangle-\alpha\langle\mathbf{g},\mathbf{g}_i\rangle \tag{5}$$

제안된 부호화기에서는 $\mathbf{r}_0$는 DC 제거된 원본 신호가 되고, LP 잔여신호는 LP 임펄스 응답과 길쌈되므로 α_i를 찾는 과정은 다음과 같이 바뀌어진다.

$$\alpha_i=\langle\mathbf{h}\otimes\mathbf{g}_i,\mathbf{r}_i\rangle \tag{6}$$

$$\begin{aligned}\mathbf{g}_i&=\arg\min_{\mathbf{g}\in\mathbf{D}}|\langle\mathbf{h}\otimes\mathbf{g},\mathbf{r}_i\rangle|\\&=\arg\min_{\mathbf{g}\in\mathbf{D}}|\langle\mathbf{h}\otimes\mathbf{g},\mathbf{r}_i\rangle-\alpha\langle\mathbf{h}\otimes\mathbf{g},\mathbf{g}_i\rangle|\end{aligned} \tag{7}$$

제안된 부호화기에서는 사전 $\mathbf{g}_i$는 정현파 사전을 사용한다.

4.2 캡스트럼(Celpstrum) LPC 잡음 부호화 방법

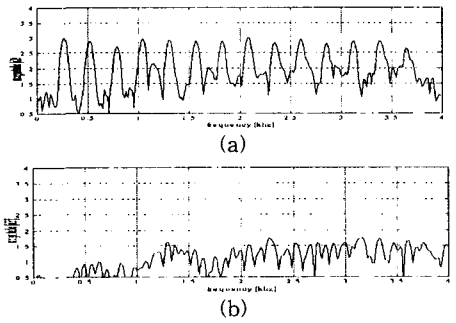
식 (8)과 같이 음성 신호$s(t)$는 여기 신호$e(t)$와 보컬트랙의 임펄스 응답$h(t)$의 길쌈과정으로 표현되며, 식 (9)와 같이 캡스트럼으로 표현될 경우 보컬트랙 성분 $h(t)$과 여기 신호성분 $e(t)$으로 분리될수 있다[5].

$$s(t)=e(t)*h(t)=(v(t)+u(t))*h(t) \tag{8}$$

$$c(t)=\mathrm{IDFT}[\log|V(w)+U(w)|+\log|H(w)|] \tag{9}$$

구체적으로, 이러한 캡스트럼$c(t)$은 큐프런시 영역에서 피치주기의 좌측 부분은 스펙트럴 포곡선을 가지는 보컬트랙 응답에 의한 성분이며, 피치 주기의 오른쪽 큐프런시 영역 부분은 여기 신호 성분으로 분류될 수 있다. 특히 피치 주기에서의 피크주변의 값은 하모닉들이 기본주파수의 배수 주변에 집중되어 있기 때문에 하모닉 성분으로 간주 될 수 있다. 따라서, 피치주기에서의 피크 주변 캡스트럼이 리프팅되고 로그 크기 스펙트럼으로 변환하여 그 음의 영역을 잡음 부분으로 정의할수 있다. 이렇게 결정된 잡음 성분들은 스펙트럴 포곡선을 LP 파라미터로 고정하는 과정을 거친다. 그 LP 파라미터는 효

율적인 양자화를 위해 LSP(Line Spectrum Pair) 파라미터로 변환된다. 변환된 LSP파라미터는 벡터양자화를 이용해 양자화 된다. 디코더에서의 합성 과정은 각 프레임 사이의 위상 일치 과정없이 가우시안 백색잡음을 LP 필터링 시킴으로써 간단히 구현된다. 제안된 부호화기에 대한 LP모델의 차수는 6차를 적용하였다. LP파라미터들과 이득 파라미터들은 4, 2비트 양자화기에의해 양자화 된다.



(a)

(b)

그림 3. 비 주기 성분의 여기신호 및 합성. (a) 원본 여기신호 스펙트럼. (b) 합성된 비 주기 신호의 스펙트럼.

4.3 IFFT를 사용한 정현파 합성법

프레임 사이의 연결성을 보장하기 위해 하모닉 파라미터들은 이전 프레임 파라미터들과 보간되어야한다. 간단한 선형 보간법이 스펙트럴 크기값과 순간 샘플링 주파수 값을 보간하기위해 사용된다. 그러나, 초기 위상 파라미터들는 프레임사이의 순간 샘플링 주파수와 샘플 크기들이 연속성을 만족하도록 보간되어야 한다. 여기서, 순간 기본 주파수는 선형으로 변한다고 가정하였다. 그러면, 합성 음성신호는 다음과 같이 표현된다[6].

$$s^k(n) = \sum_{l=1}^{L^k} A^k(l)\cos(\Delta\Phi^k(n)l + \Phi^k(l)), \qquad n = 1 \cdots N \tag{10}$$

여기서,

$$A^k(l) = \alpha(n)A_q^k(l) + (1-\alpha(n))A_q^{k-1}(l) \tag{11}$$

$$\Delta\Phi^k(n) = \sum_{m=0}^{B}(\alpha(n)w_0^k + (1-\alpha(n))w_0^{k-1}) \tag{12}$$

여기서, N은 프레임 크기이고, l은 하모닉 순열 번호, k은 현재프레임 번호, $\Phi^k(n)$은 초기 위상 값, $\Delta\Phi^k(n)$ 은 순간 위상항, $A^k(l)$은 스펙트럴 크기항, $\alpha(n) = n/N$은 선형증가함수를 나타낸다.

여기서, $s^{k-1}(N) = s^k(0)$을 만족하는 초기 위상값은 $\Phi^k(n)$ 다음과 같이 주어진다.

$$\Phi^k(l) = \frac{Nl}{2}(w_0^{k-1} + w_0^{k-2}) + \Phi^{k-1}(l) \tag{13}$$

식 (10) 을 이용한 합성 방법은 프레임 사이의 연속성을 보장하는데 성공적이지만, 복잡도가 큰 단점을 지닌다. 이러한 복잡도를 줄이기 위해 새로운 기본 파형 함수 $w(m,k)$를 다음과 같이 선언한다.

$$w(m,k) = \sum_{l=0}^{B} A_q^k(l)\cos(m\frac{2\pi}{B}l + \Phi^k(l)) \quad \text{if } l > L \text{ then } A_q^k(l) = 0 \tag{14}$$

그러면, 합성음성은 다음과 같은 식으로 정리될수 있다.

$$s^k(n) = \alpha(n)w(\frac{B}{2\pi}\Delta\Phi^k(n),k) + (1-\alpha(n))w(\frac{B}{2\pi}\Delta\Phi^k(n),k-1) \tag{15}$$

복잡도는 식(14) 에서 현저하게 감소될수 있다. 식(14)는 B가 2의 승수로 주어진다면 다음식과 같이 IFFT(Inverse Fast Fouriour Transform)으로 바뀌어지 진다.

$$\begin{aligned} w(m,k) &= \mathrm{Re}\{\sum_{l=0}^{B}\left(A_q^k(l)\cos(\Phi^k(l)) + jA_q^k(l)\sin(\Phi^k(l))\right)e^{-jm\frac{2\pi}{B}l}\} \\ &= \mathrm{Re}\{\mathrm{IFFT}\{A_q^k(jw)\angle\Phi^k(jw)\}\} \end{aligned} \tag{16}$$

식 (16)에의하여 유성음에서의 LP 여기신호는 저 복잡도로 합성될 수 있다.

5. 음악을 위한 MLT 변환 부호화기

음성을 위한 하모닉 부호화 방식은 음악신호에 대해서 부호화 할 경우 심한 왜곡 현상을 보인다. 이것은 음악신호가 단독 하모닉 구조로는 표현하기 힘들기 때문이며, 개체(Individual) 스펙트럴 라인을 사용한다 할지라도 피크 연속성 및 위상 합성에 대한 어려운 문제를 해결해야 한다. 이러한, 문제에 쉽게 해결하기 위해 완전 복원 가능 및 중첩합산(Overlapp/Add) 합성 특성 두가지를 공유한 MLT(Modulated Lapped Transform)[7]은 간단하고도 효과적인 변환 매체로 이용될 수 있다. 제안된 부호화기에서 쓰인 MLT 및 IMLT(Inverse Modualated Lapped Transform)음 식 (17), (18)과 같다.

$$M(m) = \sum_{n=0}^{N-1}\sqrt{\frac{2}{N}}\cos(\frac{\pi}{N}(n+0.5)(m+0.5))v(n) \tag{17}$$

$$\begin{aligned} v(n) &= w(N/2-1-n)x(N/2-1-n) + w(N/2+n)x(N/2+n) \\ v(n+N) &= w(N-1-n)x(N+n) + w(N/2+n)x(2N-n) \\ w(n) &= \sin\left(\frac{\pi}{2N}(n+0.5)\right) \end{aligned}$$

for $0 < n < N-1$

$$M^{-1}(n) = \sum_{m=0}^{N-1}\sqrt{\frac{2}{N}}\cos(\frac{\pi}{N}(m+0.5)(n+0.5))M(m) \tag{18}$$

$$\begin{aligned} y(n) &= w(n)M^{-1}(N/2-1-n) + w(N-1-n)M_{old}^{-1}(n) \\ y(n+N/2) &= w(N/2+n)M(n) - w(N/2-n)M_{old}^{-1}(N/2-1-n) \\ M_{old}^{-1}(n) &= M^{-1}(n+N) \end{aligned}$$

for $0 < n < N-1$

변환된 MLT 계수들은 20개의 부 대역으로 나뉘어지고, 식 (19)와 같이 LP 효과($|H(m)|$)가 가중된 $P(m)$값을 이용하여 양/음의 피크 값을 검출한다. 그러나, 이렇게 검출된 피크 값 들은 피크 크기, 피크 위치, 피크 부호등 많은 비트수를 차지하므로, 적절한 비트할당이 필요하다. 제안된 부호화기에서는 이를 위해 그림 4와 같이 각 대역당 LP 스펙트럴 응답 $|_{LP(m)}|$에 따라 피크 개수를 할당하는 방법을 사용하여, 에너지가 큰 중요 피크에 대한

비트 배당을 높이는 구조를 사용하였다.

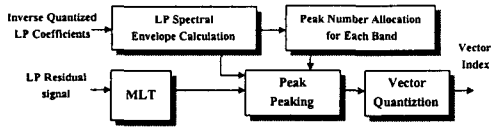


그림 4. MLT를 사용한 음악 부호화기

$$P(m)=|H(m)|M(m) \tag{19}$$

음악에서 LSP는 그 스펙트럴 변화가 적기 때문에 16 비트를 할당하였고, 제한 피크수는 10개로 하였다. 그 비트할당은 피크 크기 절대값에 20비트, 위치에 30 비트, 부호비트에 10비트를 할당하였다.

6. 실험 결과

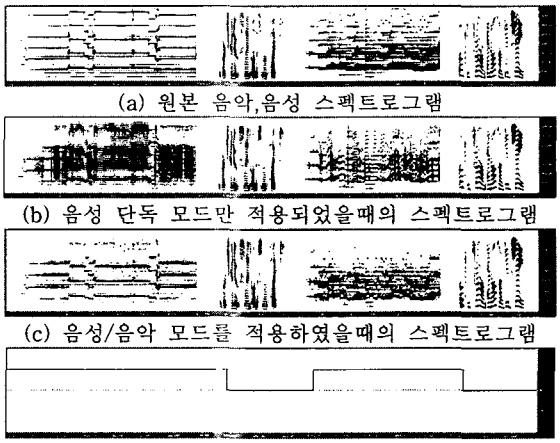
제안된 MMHTC 부호화기는 20ms 프레임 길이를 사용하며 총 40ms의 지연값을 갖는다. MMHTC의 비트할당을 표 1에 나타내었다. 제안된 부호화기를 포함한 주관적 음질 평가(MOS)를 수행하였으며, 여자 8, 남자 8, 총 16문장을 사용하였다. 비 전문가 10명으로 이루어진 청취자를 대상으로 이루어졌다. 음질 평가 결과 8kbps QCELP[8] 보다 MOS 0.2 정도 나은 음질을 얻을 수 있었다. 또한 음악부호화는 그림 5를 통해 비교할 수 있는데, 음성 단독 모드일 경우 하모닉 부호화기가 음악신호를 부호화 한다. 그 스펙트로그램에서도 알 수 있듯이 음성/음악 모드를 사용할 경우보다 원본 스펙트럴 라인에 대한 왜곡치가 큼을 알 수 있다.

표 1. 4kbps MMHTC 부호화기의 비트할당

파라메터	유성음	무성음	음악
LSP	24		16
V/UV/M	2		
Pitch	7	0	0
Magnitudes Gain	5	0	0
Magnitudes Shape	36	0	0
Noise Gain/Shape	6	0	0
Time Domain Shape	0	54	0
MLT Asolute Mgnitudes	0	0	20
MLT sign	0	0	10
MLT position	0	0	30
Total	48/20ms		

7. 결론

본 논문에서 음성 및 음악에 대해 다중모드를 갖는 향상된 하모닉 잡음 부호화 방식과 MLT 변환 부호화 방식을 제안한다. 하모닉 부호화 방법은 MP(Matching Pursuit)을 이용한 스펙트럴 크기 분석법 및 캡스트럼 LPC 잡음 스펙트럴 예측기, 빠른 정현파 합성법을 제안하고 있으며, 음악에 대해서는 LP 적응 피크검출을 통해 저 전송률에서 비트수 제한 대한 표현 능력의 단점을 보완하였다. 제안된 4kbps MMHTC는 음질 시험결과 8KQCELP 보다 좋은 음질을 보여주었다.



(a) 원본 음악,음성 스펙트로그램

(b) 음성 단독 모드만 적용되었을때의 스펙트로그램

(c) 음성/음악 모드를 적용하였을때의 스펙트로그램

(d) 음성(0) 음악(1) 분류기의 출력

그림 5. MMHTC의 실험결과

※ 본 연구는 2000년도 정보통신부 대학기초연구 지원사업(과제번호 2000-085-02)의 지원으로 수행되었습니다.

참고문헌

[1] T. Moriya, N. Iwakami, A. Jin, K. Ikeda, and S. Miki,"A Design of Transform Coder for Both Speech and Audio Signals at 1 bit/samples," in *Proc. IEEE Int. Conf. Acount., Speech, Signal Processing*, pp. 1371-1374, 1997.

[2] S. A. Ramprashad, "A Two Stage Hybrid Embedded Speech/Audio Coding Structure," in *Proc. IEEE Int. Conf. Acount., Speech, Signal Processing*, pp. 337-340, 1998.

[3] ISO/IEC JTC1/SC29/WG11, " Information Technology- Coding of Audiovisual Objects Part 3: Audio Subpart2: Parametric Coding." N1903PAR, 1997

[4] M. Goodwin, *Adaptive Signal Models*, Kluwer Academic Publishers, 1998.

[5] B. Yegnanarayana, Christophe dAlessandro and Vassilis Darsinos, "An Iterative Algorithm for Decomposition of Speech Signals into Periodic and Aperiodic Components," *IEEE Transaction on speech and audio processing*, vol. 6, NO. 1,pp. 1-11,1998.

[6] R. J. McAulay, T. F. Quartieri, Sinusoidal coding, *Speech Coding and Synthesis*, Chapter 4, W. B. Kleijn, and K. K. Paliwell Eds., Elsevier, 1995.

[7] H. Malvar "Fast Algorithms for Orthogonal and Biothogonal Modulated Lapped Transforms," in *Proc IEEE Symposium, Advances in Digital Filtering and Signal Processing*, 159-163, 1998.

[8] P. J. A. DeJaco, W. Gardner and C. Lee, "QCELP: North American CDMA digital cellular variable rate speech coding standard," in *Proc. IEEE Workshop on speech Coding for Telecommunications*, (Sainte-Adele. Quebec), pp. 5-6, 1993.