2000年度 第13回 信號處理合同學術大會 論文集 第13卷 1號

# 정현파 모델을 이용한 2.4kbps 음성부호화 알고리즘

백 성 기, 배 건 성
경북대학교 전자·전기 공학부

# 2.4kbps Speech Coding Algorithm Using the Sinusoidal Model

Sung Gi Baek, Keun Sung Bae
School of the Electronic & Electrical Engineering, Kyungpook National University

sung@mmirl1.knu.ac.kr, ksbae@ee.knu.ac.kr

## 요 약 문

STC(Sinusoidal Transform Coding) 방식은 음성신호의 주파수 영역에서 스펙트럼 피크치들을 정현파로 모델링하여 합성하는 방식을 말한다. 저전송률 STC 방식에서는 전송되는 정보량을 줄이기 위해 스펙트럼 피크를 대신해 음성신호의 스펙트럼 포락선 정보와, 피치정보를 이용하여 얻어지는 고조파 성분들을 정현파로 모델링하여 음성을 합성한다. 본 논문에서는 음성신호의 정현파 모델에 기반하여 2.4kbps 전송속도를 갖는 음성부호화 알고리즘을 제안하였으며, 실험결과로 합성음의 파형과 스펙트럼 특성, 위상특성, 그리고 MOS(Mean Opinion Score) 테스트를 이용한 합성음의 음질을 비교/분석하였다.

## I. 서 론

STC 방식은 정현파 모델을 이용한 음성부호화 방식으로, 주파수 영역에서 스펙트럼의 피크치들을 정현파로 모델링하여 합성하는 방식을 말한다[1]. 정현파 모델에 기반한 저전송률 음성부호화기에서는 전송되는 정보량을 줄이기 위해 스펙트럼 피크 대신에 스펙트럼 포락선 정보와 피치정보를 이용한다. 즉, 유성음인 경우에는 스펙트럼의 기본주파수와 고조파에 해당하는 성분분들을 정현파로 표현하고, 여기신호가 무성음인 경우에는 주기성분이 일정하지 않으므로 주파수 스펙트럼에서 일정 간격으로 크기를 찾아서 그에 해당하는 정현파를 생성하게 된다.

정현파 모델에서, 각 분석 프레임의 합성음 $s(n)$은 식 (1)과 같이 표현할 수 있다.

$$s(n) = \sum_{l=1}^{L} A_l \cos(w_l n + \phi_l) \qquad (1)$$

여기에서 $A_l$과 $\phi_l$은 주파수 $w_l$에 해당하는 스펙트럼의 크기 및 위상을, $L$은 합성에 사용되는 정현파의 수를 나타낸다. 식 (1)과 같이 정현파 모델을 이용하여 음성을 합성하기 위해서는 음성신호의 STFT(Short Time Fourier Transform)의 피크치, $A_l$을 필요로 한다. 그러나 저전송률 음성부호화 알고리즘에서는 모든 피크치를 전송할 수 없으므로, 피치정보를 이용한 기본주파수와 고조파, 그리고 스펙트럼 포락선 정보를 이용하여 음성신호의 스펙트럼 피크, $A_l$을 대신한다. 따라서 정현파 모델에 기반한 저전송률 음성부호화기에서는 고조파성분을 위한 피치정보, 그리고 위상정보와 스펙트럼 크기를 위한 스펙트럼 포락선정보를 필요로 한다.

본 논문에서는 정현파 모델을 이용한 2.4kbps 음성부호화 알고리즘을 제안하고, 합성음의 파형과 스펙트럼 특성, 그리고 MOS 테스트를 이용한 합성음의 음질을 비교/분석하였다. 제안한 알고리즘은 25ms 분석프레임마다 음성신호를 분석하여 정현파 모델 파라미터를 추출하여 부호화하며, 피치정보, 스펙트럼 크기정보, 위상정보를 전송파라미터로 한다. 스펙트럼 크기정보는 스펙트럼 포락선을 이용하는데, SEEVOC (Spectral Envelope Estimation VOCoder)[2] 알고리즘과 LPC[3]를 이용하여 추정하는 기법을 이용하였다[4].

본 논문의 구성은 다음과 같다. 2장에서는 제안한 정현파 모델을 이용한 2.4 kbps 음성부호화 알고리즘에 대해서 설명한다. 3장에서는 음성부호화 알고리즘에 의해 얻어진 합성음을 비교/분석하며, 마지막으로 4장에서 결론을 맺는다.

## Ⅱ. 정현파 모델을 이용한 2.4 kbps 음성 부호화 알고리즘

### 1. 정현파 모델을 이용한 음성부호화 알고리즘

8kHz, 16bits로 샘플링 되고 양자화된 25ms의 분석 프레임을 갖는 입력음성은 전처리과정을 거친 후, 유/무성음의 이원적인 모델로 나뉘어져 처리되고, 12.5ms의 overlap을 갖는다. Encoder에서는 피치, 스펙트럼 포락선을 위한 LPC 계수와 이득, 추정한 위상정보가 양자화된 후 전송되며, decoder에서는 전송된 파라미터를 이용하여 식 (2)와 같이 음성을 합성한다.

$$s(n)=\sum_{l=1}^{N} A_l \cos(lw_o n+\phi_l) \qquad (2)$$

여기에서 $A_l$과 $\phi_l$은 기본주파수 $w_o$와 고조파 $lw_o$에 해당하는 스펙트럼의 크기 및 위상을, $L$은 합성에 사용되는 정현파의 수를 나타낸다. 그림 1은 encoder와 decoder의 블럭도를 나타낸 것이다

입력음성은 140 Hz의 cutoff 주파수를 갖는 high-pass filter와 down-scaling으로 이루어진 전처리 과정을 거친 후 처리된다. 식 (3)은 high-pass filter와 2배의 down-scaling에 사용된 필터의 전달함수이다.

$$H(z)=\frac{0.46363718-0.92724705z^{-1}+0.46363718z^{-2}}{1-0.9059465z^{-1}+0.9114024z^{-2}} \qquad (3)$$

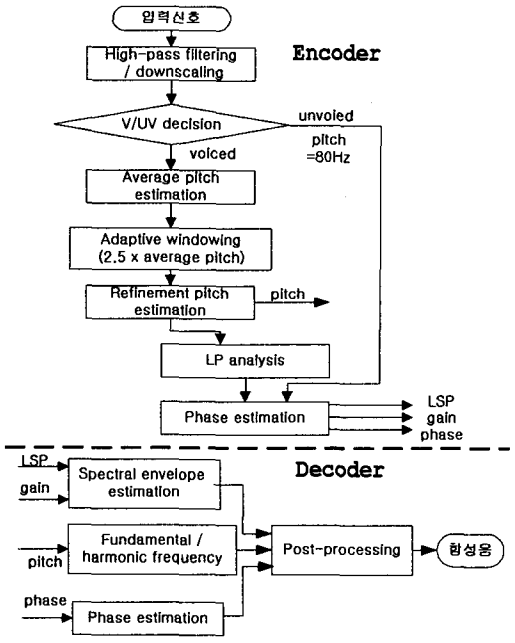


그림 1. 정현파 모델을 이용한 음성부호화 알고리즘의 encoder/decoder 블럭도

전처리과정을 거친 입력신호는 zero-crossing rate와 에너지를 이용하여 유성음과 무성음으로 구별된다. 유성음의 경우엔 피치정보와 위성정보, 스펙트럼정보를 찾으며, 무성음은 스펙트럼 정보만을 찾고 위상정보로는 80Hz 간격으로 $-\pi \sim \pi$ 범위에서 uniformly distributed random number를 사용한다. 유성음의 경우 AMDF (Average Magnitude Difference Function)알고리즘[5]을 이용하여 구한 대략적인 피치와 이전프레임의 피치를 이용하여 평균피치를 구한다. 평균피치의 2.5배에 해당하는 길이를 adaptive window 길이로 정하며[6], 전처리과정을 거친 음성에 adaptive window를 적용하여 정확한 피치와 스펙트럼 포락선, 위상정보를 구하게 된다. 그림 2는 분석프레임과 adaptive window를 나타낸 것이다.

정확한 피치정보는 모든 스펙트럼의 피크에 해당하는 주파수를 이용한 합성음과 피치에 의한 기본주파수와 고조파를 이용한 합성음과의 에러가 최소가 되도록 주파수 영역에서 피치를 찾는다. 정현파 모델에 기반한 저전송률 음성부호화기는 스펙트럼 포락선을 이용하여, 기본주파수와 고조파에 해당하는 스펙트럼 크기를 얻는다. 따라서 스펙트럼 포락선은 기본주파수와 고조파에 해당하는 스펙트럼 피크를 잘 따라가야 된다. 본 논문에서는 SEEVOC 알고리즘을 사용하여 스펙트럼의 모든 피크 중에서 기본주파수와 고조파에 해당하는 특정 피크를 검출한 후, 이 피크를 사용하여 14차 LPC 계수를 구하는 방법을 이용하였으며, 구해진 LPC 계수를 이용하여 스펙트럼 포락선을 얻는다[4]. 위상정보는 여기신호 pulse의 시작 시기를 나타내는 onset time과 minimum phase system을 이용한 vocal tract의 위상을 이용하여 추정하였다[8]. 위상정보는 식(4)와 같이 표시된다.

$$\hat{\phi}(w)=-\hat{n_o}w+\Phi_s(w)+\beta\pi \qquad (4)$$

여기서 $\hat{\phi}(w)$는 추정한 위상정보이고, $\hat{n_o}$는 onset time, $\Phi_s(w)$는 vocal tract의 위상을 나타낸다. Minimum phase system의 경우, 합성음 $s(n)$과 $-s(n)$의 위상이 구분되지 않으므로 $\beta$로 하여금 ±부호를 결정하도록 한다.

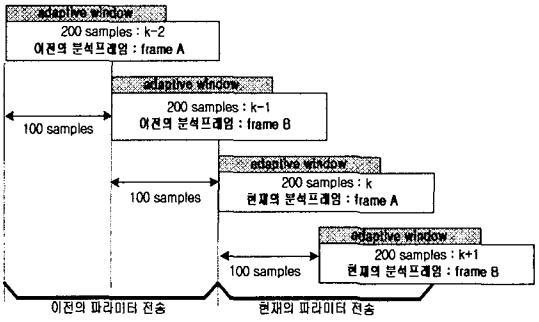


그림 2. 분석프레임 및 adaptive window

Encoder에서 구해진 유/무성음 정보, 피치, LPC 계수, 이득, 위상정보는 60bits로 양자화되어 전송된다. Decoder에서는 전송된 파라미터를 이용하여 합성음을 구한 후, 후처리 과정을 거친다. 후처리는 100Hz의 cutoff 주파수를 갖는 high-pass filter와 전처리의 down-scaling을 보상하기 위한 up-scaling으로 이루어져 있다. 식 (5)는 사용된 high-pass filter이다.

$$H(z)=\frac{0.93980581-1.8795834z^{-1}+0.93980581z^{-2}}{1-1.9330735z^{-1}+0.93589199z^{-2}} \quad (5)$$

합성음은 프레임간 연속성을 위하여 overlap-and-add 합성 방식을 수행하여 최종적인 합성신호를 생성하게 된다.

2. 전송파라미터의 전송과 2.4 kbps 양자화 기법

전송되는 파라미터는 유성음일 경우엔 유성음 정보, 피치정보, 위상정보( $\widehat{n_0}, \beta$ ), 스펙트럼 포락선 정보(LPC 계수), 이득으로 구성되며, 무성음일 경우엔 무성음 정보, 스펙트럼 포락선 정보, 이득이며, 피치와 위상정보는 아무런 의미가 없는 0을 전송한다. 파라미터의 전송은 200samples 마다 이루어지며, 두 개의 분석프레임을 갖는다. 첫 번째 분석프레임을 frame A, 두 번째 분석프레임을 frame B라 하며, 구성은 그림 2와 같다.

표 1은 2.4kbps STC의 비트할당을 나타낸 것이다. 전송파라미터의 양자화 방법은 다음과 같다. 유/무성음 정보는 1bit를 할당하여 0 또는 1을 전송한다. Fame B의 LPC 계수는 유/무성음에 따라 LSF(Line Spectral Frequency)로 바꾸어 (6, 8)의 multistage split VQ를 이용하여 각각 8, 6, 6bits가 할당되며, decoder에서 frame A의 LSF는 이전분석프레임:frame B의 LSF와 현재분석프레임:frame B의 LSF를 식 (6)과 같이 보간하여 사용한다.

$$Frame\ A: LSF_i^{(A)} = 0.5LSF_i^{(previous\ B)} + 0.5LSF_i^{(current\ B)}$$
$$Frame\ B: LSF_i^{(B)} = LSF_i^{(current\ B)}, \quad i=1,\ldots,14 \quad (6)$$

본 실험에서 사용되는 피치는 시간영역에서의 피치의 역수를 말한다. Frame B의 피치는 그림 3과 같이 8bits가 할당되며, 대수적으로 양자화된다. Frame A에서의 피치는 이전분석프레임:frame B의 피치와 현재분석프레임:frame B의 피치를 이용하여 식 (7)의 후보피치를 구한 후, 에러가 적은 피치의 인덱스를 2bits로 양자화하여 전송한다.

$$\begin{aligned} pitch_k^0 &= pitch_{k-1} \\ pitch_k^1 &= pitch_{k+1} \\ pitch_k^2 &= 0.5\times pitch_{k-1} + 0.5\times pitch_{k+1} \\ pitch_k^3 &= 0.25\times pitch_{k-1} + 0.75\times pitch_{k+1} \end{aligned} \quad (7)$$

표 1. 2.4kbps STC의 비트 할당

| | subfame 1 | subframe 2 | Total |
|---|---|---|---|
| 유/무성음 | 1 | 1 | 2 |
| L S P | | 20 | 20 |
| 피 치 | 2 | 8 | 10 |
| onset time | 8 | 8 | 16 |
| $\beta$ | 1 | 1 | 2 |
| 이 득 | 5 | 5 | 10 |
| Total | | | 60 |

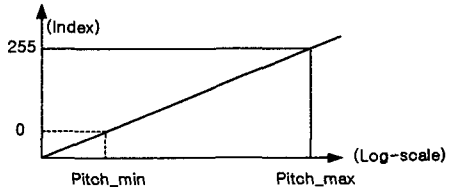


그림 3. 피치의 양자화 비트 할당

위상정보로는 onset time( $\widehat{n_0}$ )과 $\beta$가 있으며, frame A, B에 대해서 같은 양자화 비트가 할당되고, 스칼라 양자화된다. 이득은 frame A, B에 대해서 각각 6bits가 할당되며, Lloyd-Max 알고리즘을 이용하여 비선형적으로 양자화 된다.

## Ⅲ. 실험 및 고찰

본 연구에서는 8kHz로 샘플링되고, 16bits로 양자화된 임의의 음성데이터를 이용하여 원음성과 합성음의 파형, 스펙트럼, 위상을 비교/분석하였으며, 합성음의 음질평가를 위해서 주간적인 음질평가 방법인 MOS 테스트를 이용하였다. 그림 4는 원음성과 정현파모델에 기반한 2.4kbps 음성부호화기를 이용한 합성음의 파형을 나타낸 것이며, 그림 5와 그림 6은 원음성과 합성음의 스펙트럼과 위상을 나타낸다. STC 방식이 파형부호화 방식이 아니므로 그림 4의 합성음의 파형은 원음성에 비해 약간의 왜곡이 발생한다. 그러나, 그림 5의 스펙트럼 특성에서는 원음성의 포만트 정보를 대부분 가지고 있으며, 그림 6의 위상에서는 원음성의 위상과 합성음의 위상이 거의 일치함을 알 수 있다.

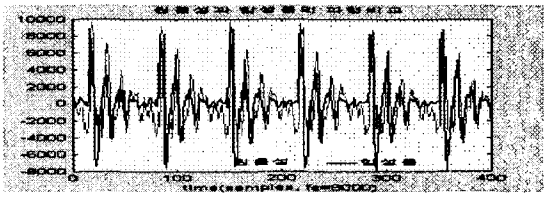

그림 4. 원음성과 합성음의 파형 비교

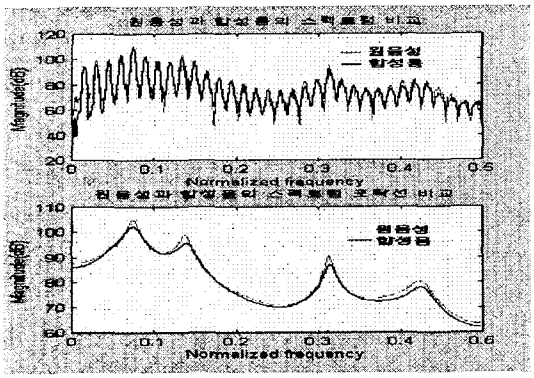


그림 5. 원음성과 합성음의 스펙트럼 비교

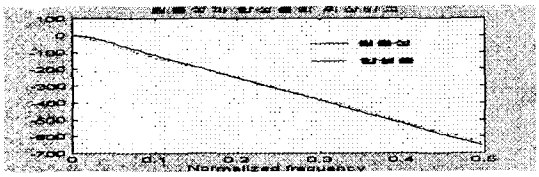

그림 6. 원음성과 합성음의 위상 비교

표 2는 4.8kbps DoD CELP 음성부호화 방식과 제안한 정현파 모델을 이용한 2.4kbps 음성부호화 알고리즘에 대해 10명의 청취자를 대상으로 합성음의 음질을 주관적인 음질평가 방법인 MOS 테스트를 이용하여 informal한 청취실험을 수행한 결과를 보인 것이다. 합성실험에 사용된 음성은 남성화자가 발성한 아래 3개의 문장을 이용하였다. 4.8kbps CELP 음성부호화 방식보다는 음질이 전체적으로 약간 떨어지지만, 대체적으로 MOS 3.5 이상의 합성음을 얻을 수 있음을 알 수 있다.

문장 1 : 하늘을 날고자 하는 인간의 욕망은 끝이 없습니다.
문장 2 : 한국의 가을하늘은 참으로 맑고 푸르릅니다.
문장 3 : We saw the ten pink fishes.

표 2. 합성음의 MOS 테스트

| | 문장 1 | 문장 2 | 문장 3 | Total |
|---|---|---|---|---|
| 2.4kbps STC | 3.5 | 3.9 | 4.0 | 3.8 |
| 4.8kbps CELP | 3.5 | 4.2 | 4.3 | 4.0 |

## Ⅳ. 결 론

STC 방식은 정현파 모델을 이용한 음성부호화 방식으로, 주파수영역에서 음성신호의 스펙트럼 피크 성분들을 정현파로 모델링하여 합성하는 방식을 말하며, 일반적으로 저전송률에서는 음성신호의 고조파 성분들을 정현파로 모델링하여 합성한다.

본 논문에서는 정현파 모델을 이용한 2.4kbps 음성부호화 알고리즘을 제안하고, 합성음의 파형과 스펙트럼 특성, 그리고 MOS 테스트를 이용한 합성음의 음질을 비교/분석하였다. 남성화자의 음성에 대한 실험결과, 합성음 파형에서는 원음성에 비해 약간의 왜곡이 발생하지만, 스펙트럼 특성에서는 합성음의 스펙트럼 특성이 원음성의 포만트 정보를 대부분 가지고 있으며, 위상정보도 원음성의 위상을 잘 따라가는 것을 볼 수 있다. 또한, 합성음의 음질테스트에서는 대체적으로 MOS 3.5 이상의 음질을 가짐을 확인하였다. 그러나, 여성화자의 음성은 정확한 피치정보의 추정이 잘 안되고 double 피치가 발생하여 음질의 저하를 초래하였다. 앞으로 정확한 피치추정기법을 통한 여성화자의 음질 개선에 대해 연구하고자 한다.

본 연구는 1999년도 정보통신부 대학기초연구지원사업의 지원으로 수행되었으며, 지원에 감사드립니다.

## Ⅴ. 참 고 문 헌

[1] A.S.Spaniass,"Speech Coding : A Tutorial Review", Proc. of IEEE, Vol.82, No.10, Oct., 1994.
[2] Paul, D.B., "The specral envelope estimation vocoder", IEEE Trans.,Acoustic, Speech, Signal Processing., ASSP-29:pp786-794, 1981
[3] A.El-Jaroudi and J.Makhoul, "Discrete all pole modeling", IEEE Trans. Acoust., Speech, Signal Processing, 39(2):pp.411-423, Fev 1991.
[4] 백성기, 배건성, '음성신호의 정현파 모델에 기반한 저전송률 부호화를 위한 스펙트럼 포락선 추정 기법에 관한 연구', 제17회 음성통신 및 신호처리 학술대회, 17권1호, pp. 197-200, 2000.8
[5] A.M.Kondoz, "Digital Speech", pp.61-63
[6] R.J.McAulay and T.F.Quatieri, "Low-rate speech coding based on the sinusoidal model", Advances in Speech Signal Processing, chapter6, pp.165-208, Marcel Dekker, 1993
[7] R.J.McAulay and T.F.Quatieri, "Pitch estimation and voicing detection based on a sinusoidal model", Proc. IEEE, Accoustics, Speech, and Signal Processing, 1990
[8]. R.J. McAulay, T.F. Quatieri, 'Sine-wave phase coding at low data rates', Acoustics, Speech, and Signal Processing, ICASSP-91. 1991