TTS DB 압축을 위한 광대역 파형보간 부호기 구현

양희식(ICU), 한민수(ICU)

<차 례>

1. 서론
2. 기존의 음성 압축알고리즘 검토
  2.1. 개요
  2.2. 기존의 음성 부호기
  2.2. 기존의 음성 부호기들의 성능
3. 파형 보간 음성압축 알고리즘
  3.1. 개요
  3.2. 파형보간 부호기 개요
  3.3. 파형보간 부호화기의 구조
  3.4. 파형보간 복호화기의 구조
  3.5. 양자화
4. WI 부호기 구현
5. 결론

<Abstract>

Implementation of Wideband Waveform Interpolation Coder for TTS DB Compression

Heesik Yang, Minsoo Hahn

The adequate compression algorithm is essential to achieve high quality embedded TTS system. In this paper, we propose waveform interpolation coder for TTS corpus compression after many speech coder investigation. Unlike speech coders in communication system, compression rate and quality are more important factors in TTS DB compression than other performance criteria. Thus we select waveform interpolation algorithm because it provides good speech quality under high compression rate at the cost of complexity. The implemented coder has bit rate 6kbps with quality degradation 0.47. The performance indicates that the waveform interpolation is adequate for TTS DB compression with some further study.

* Keywords: Waveform interpolation, TTS DB compression, Speech coder, Characteristic waveform.

1. 서 론

기존의 PC 기반의 음성합성 시스템은 그 특성상 단독으로 상품화되기 힘들다. 그런 이유로 국내외 음성 합성 시장은 그 입지가 좁아져 가고 있는 상황이다. 그러나 최근 각광 받고 있는 휴대용 단말기에서 합성 시스템의 효용성은 그 가능성이 무한하고 음성 합성 시장의 활로를 열어 줄 것으로 기대된다. 따라서 휴대용 단말기에 탑재할 수 있는 고품질 소용량 음성합성기 개발은 음성 합성 연구자들이 시급히 해결해야 할 과제 중 하나로 대두되고 있다.

합성 시스템의 용량을 줄이는 방법은 합성 시스템 초창기의 규칙기반 합성 시스템이나 대표 단위(unit)에 신호 처리를 이용하여 접합하는 방법을 고려해 볼 수 있다. 그러나 이러한 합성 시스템들은 현재 코퍼스 기반 합성 시스템에 비해 음질이 너무 낮다는 것은 널리 알려진 사실이다. 따라서 현재의 합성 음질을 유지한 채 합성기의 용량을 줄이는 방법이 깊이 연구되어야 하며 합성기 용량의 대부분을 음성 DB가 차지하고 있는 만큼 DB의 용량을 줄이는 것이 올바른 접근 방법일 것이다. 본 논문에서는 DB 용량을 줄이려는 시도 중의 하나로, 단위 삭제(pruning)를 배제하고 음성 DB 자체가 가지고 있는 잉여 정보를 제거하고 음성 특성을 반영한 압축 알고리즘을 이용하여 음성 DB를 부호화함으로써 합성 DB의 용량을 최대한 줄임과 동시에 음질 저하를 최소화 하고자 한다. 이를 위하여 제2장에서는 기존의 음성 압축 알고리즘을 검토하고 제3장에서는 합성 DB 압축에 가장 우수한 성능이 기대되는 파형보간(WI: waveform interpolation) 음성압축 알고리즘을 설명한다. 제 4장에서는 3장에서 설명한 WI 부호기의 구현 및 성능 평가 결과에 대해서 언급하고 마지막으로 5장에서 결론을 맺도록 한다.

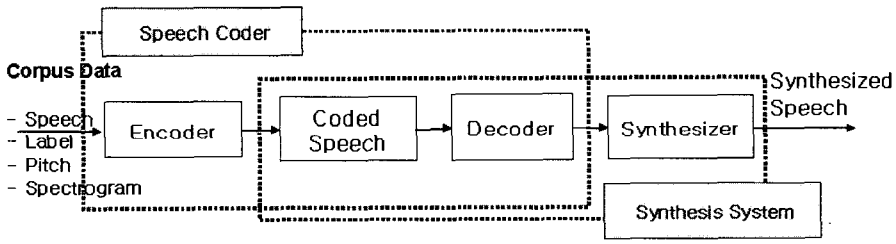
2. 기존 음성 압축알고리즘 검토

2.1. 개요

음성 코딩이란 디지털 포맷의 음성 입력 신호를 부호기에서 저 용량 비트스트림으로 변환하고 복호기에서 부호화된 비트스트림으로부터 원래 음성을 복원하는 기술로 정의된다[9]. 따라서 합성 시스템의 대용량 코퍼스를 음성 코딩한다는 것은 저 용량의 비트스트림으로 음성 DB를 변환하는 것이 된다. 일반적인 음성 부호기는 bit rate, 복원 음질, 연산복잡도, 지연 시간으로 성능을 평가한다. 먼저 bit rate은 bits/s으로 표현되며 원 신호에 대한 압축률 산출의 근거가 된다. 복원한 음성의 음질은 일반적으로 주관적 음질 평가를 실시하며 주로 쓰이는 3가지 방법은 diagnostic rhyme test(DRT), diagnostic acceptability test(DAM), mean opinion

score(MOS)가 있다. 연산복잡도는 일반적으로 MIPS로 측정하며 연산 복잡도가 증가하면 하드웨어 전력 소모가 증가하고 또한 실시간 구현이 어렵게 된다. 지연시간은 입력신호가 부호기와 복호기를 거쳐 복원될 때까지의 소요 시간을 의미하며 일반적인 대화형 통신 환경에서는 150ms, 단 방향 통신의 경우 400-500ms 이상이 되면 음질에 영향을 준다[9].

합성 DB 압축에 사용되는 음성 부호기의 경우 일반적인 부호기와는 다른 특성을 갖는다. 즉, <그림 1>에서 부호화 과정과 복호화 과정이 실시간으로 이루어질 필요가 없으며 복호화 과정만을 고려하면 된다. 따라서 일반적인 음성 부호기와는 달리 연산복잡도와 지연시간은 상대적으로 음성 DB 압축 부호기의 성능에 적은 영향을 끼치게 된다. 이와 같은 이유로 합성 DB 압축에 사용되는 부호기는 높은 압축률과 양질의 복원 음질을 가지는 것에 초점을 맞추어 알고리즘을 개발할 수 있다.



<그림 1> 음성압축 시스템 개요도

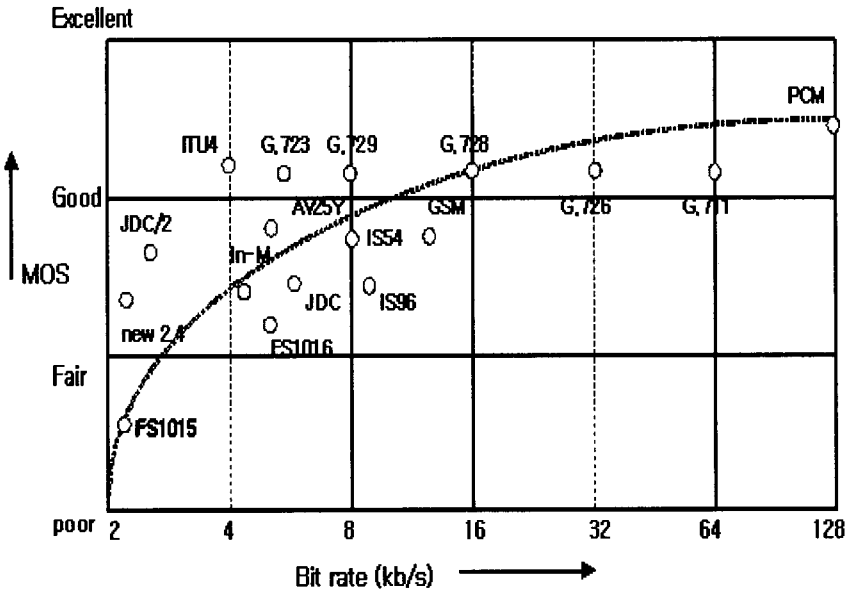
2.2. 기존의 음성 부호기

음성 부호기는 일반적으로 파형(waveform) 코딩방식과 파라미터(parametric) 또는 음원(source)코딩 방식으로 분류하지만, 최근의 hybrid 방식 부호기들은 이 기준으로 분류하기 어렵다. 복원 신호가 원 신호로 수렴 여부에 따라 파형근사(waveform approximating)와 파라미터(parametric) 부호기로 분류하면[9], 파형근사 부호기는 예측(predictive)코딩, 부대역(subband) 코딩으로 분류되고 ADPCM을 비롯한 multi-pulse, regular-pulse, CELP(code-excited linear prediction)와 같은 hybrid 부호기들도 예측코딩에 속한다. 파라미터 코딩은 선형예측기반 보코더, 정현파 부호기, 파형보간 부호기로 분류된다.

2.3. 기존 부호기들의 성능

일반적으로 파형근사 (waveform approximating) 부호기의 경우 음질은 우수하나 압축률에 한계가 있으며, 파라미터방식(parametric) 부호기의 경우 압축률은 높으나

좋은 음질을 구현하기 힘들다. <그림 2>는 현재 표준화되어 있는 부호기들의 성능을 보여준다[7][9][12]. 현재 CELP 기반의 hybrid 부호기가 음질대비 높은 압축률을 가져 표준 부호기로 널리 사용되고 있으나 이러한 파형근사 부호기는 4 kb/s 이하에서 음질 저하가 급격히 일어나 높은 압축률을 요구하는 합성 DB의 압축에는 사용되기 어렵다.



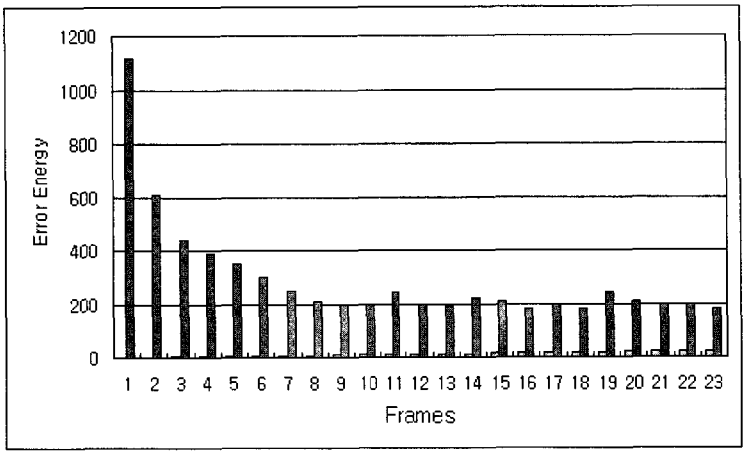
<그림 2> 표준화 부호기들의 압축률 대비 음질

3. 파형보간 음성 압축알고리즘

3.1. 개요

최근 표준화된 부호기들은 CELP기반 혼합형(hybrid) 방식의 부호기가 주류를 이룬다. 이는 CELP 기반 부호기가 높은 압축률 대비 양질의 복원 음질을 가지기 때문인데 합성DB의 압축에도 CELP에 기반한 여러 시도가 있었다. 그러나 음성 DB 압축은 일반적인 통신 환경에서의 부호기와 비교할 때, 더욱 높은 수준의 복원 음질을 요구하며 압축률 또한 높아야 한다. 부가적으로 일반적인 부호기의 경우, 연속된 입력 음성을 부호화 하고 부호화된 비트스트림으로부터 연속된 음성 렬(string)을 복호화하게 되지만 합성 시스템에 탑재될 음성 부호기의 경우 복호화 과정은 연속된 음성 렬(string)의 복호화가 아니라 합성 시스템에서 선택된 개별 합

성 단위의 복호화 과정 후 접합(concatenation)을 통해 합성음을 재생하게 된다. 따라서 분석 창의 크기에 따른 부가적인 오버헤드가 존재하며 이것은 압축률의 현저한 저하를 유발한다. 두 번째로 복호화 과정에서 발생한 초기 오차는 <그림 3>와 같은 특성을 가진다[8]. 그러므로 접합 단위가 짧을 경우 초기 오차가 제거되기 전에 다음 단위의 복호화가 이루어지므로 일반적인 부호기에서 보다 음질 열화가 더욱 크게 일어난다. 초기 오차의 전파 및 개별 단위의 복호화로 인한 오버헤드는 합성 DB 압축에 CELP 기반 부호기들의 사용을 제한하며 따라서 CELP 기반 부호기보다 상대적으로 높은 압축률을 가지며 양질의 복원음질을 유지하는 알고리즘의 개발을 요구하게 된다. 그러므로 우리는 높은 압축률에도 양질의 복원음질을 가지지만 많은 연산량으로 인해 일반적인 통신 부호기로는 큰 주목을 받지 못했던 파형보간(waveform interpolation) 부호기를 채택하였다.



<그림 3> 음성 부호기에서의 오차 전파 패턴

3.2. 파형보간(waveform Interpolation) 부호기 개요

파형보간 부호기는 유성음 구간에서 파형을 피치 주기로 나누면 피치 주기 간 신호의 변화는 크지 않으므로 피치 주기의 신호 일부만을 부호화 하고 그 사이의 신호는 선형 보간을 이용하여 복원할 수 있다는 점을 이용한다. 초기의 파형 보간 기법은 유성음의 특성을 이용하기 때문에 다른 음성 코딩 방식에 보조적으로 쓰였으나[10] 이후 음성 신호를 식 (1)과 같이 SEW(slowly evolving waveform), s_n와 REW(rapidly evolving waveform), r_n의 결합으로 정의하고 식 (2)에서와 같이 저역 필터링(low-pass filtering)을 통해 해체, 각각의 신호가 다른 코딩 해상도와 양자화 해상도를 가지도록 효과적인 압축이 가능하다는 특성 파형(characteristic waveform)

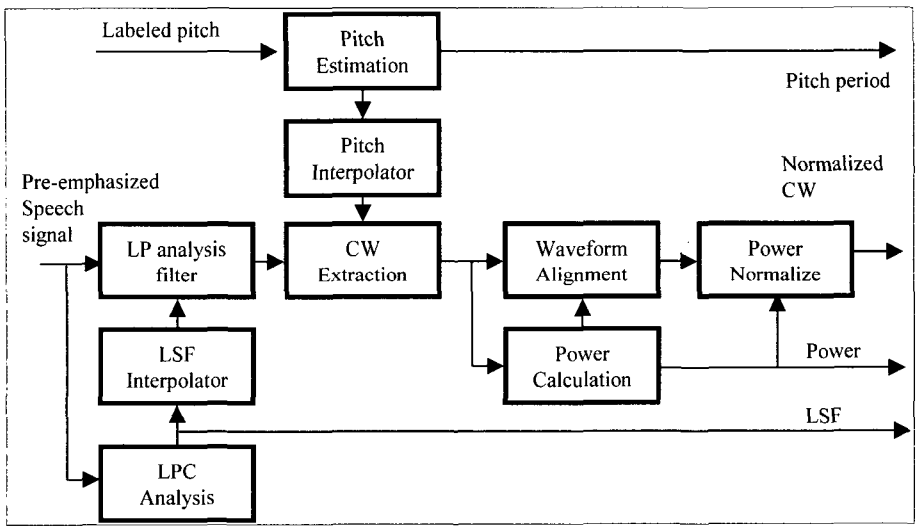
개념이 도입되면서 파형 보간 기법은 단독으로 쓰일 수 있게 되었다.

$$c_n = s_n + r_n \tag{1}$$

$$s_{n-[N/2]} = \sum_{i=0}^{N-1} a_i c_{n-i}, \qquad r_n = c_n - s_n \tag{2}$$

3.3. 파형보간 부호화기의 구조

입력된 음성 신호는 <그림 4>의 부호화 과정에 따라 부호화되며 각각의 과정은 다음과 같다.



<그림 4> 파형보간 부호기 블록 다이어그램

3.3.1. 선형예측(Linear Prediction) 분석

입력신호는 pre-emphasis하여 고주파 에너지를 보상하고 선형예측 분석하여 선형예측 계수를 구한다. 인접하는 프레임 간의 선형예측계수의 급격한 변화를 피하기 위하여 구해진 계수를 LSF 계수로 변환하고 LSF 도메인에서 부(sub)프레임 단위로 선형보간한다. 선형보간된 LSF를 다시 선형예측 계수로 변환하고 입력신호의 선형 예측 필터링을 통해 잔차신호를 구한다[3].

3.3.2. 특성 파형(Characteristic waveform) 추출

특성 파형은 분석 구간에서 파형 추출 비율에 따라 정해진 추출 지점을 일정하게 설정하고 주어진 지점에서 선형 보간된 피치 길이를 가지는 파형을 DTFS(discrete time Fourier spectrum)로 나타낸다. 따라서 식 (3)의 $s(m)$은 개별파형 추출지점에서의 특성 파형을 의미하며 A_k와 B_k는 DTFS 계수, p는 추출 지점에서의 피치 제곱을 나타낸다. 식(4)의 $s(m,n)$은 개별 특성파형에 시간 차원을 더하여 1차원의 잔차신호를 2차원의 특성 파형으로 표현하게 된다.

$$s(m) = \sum_{k=0}^{[P/2]} \left[A_k \cos\left(\frac{2\pi km}{P}\right) + B_k \sin\left(\frac{2\pi km}{P}\right) \right] \qquad 0 \le m < P \qquad (3)$$

$$s(n,m) = \sum_{k=0}^{[P(n)/2]} \left[A_k(n) \cos\left(\frac{2\pi km}{P(n)}\right) + B_k(n) \sin\left(\frac{2\pi km}{P(n)}\right) \right] \quad 0 \le m < P(n) \quad (4)$$

3.3.3. 특성 파형 정렬

연속된 특성 파형의 피크 성분이 동일 위상(phase)을 가지도록 인접하는 특성파형을 정렬함으로서 피치 주기의 변화를 쉽게 얻을 수 있도록 하는 것이 목적이며, 파형의 정렬은 세 가지 시나리오를 따른다. 첫째 인접하는 파형의 길이가 동일 할 경우, 둘째 인접하는 파형에서 앞선 파형의 길이가 긴 경우, 마지막으로 두 번째와 반대의 경우 세 가지에 따라 나뉘며 주파수 영역에서 영 삽입(zero padding)과 circular shift를 이용 정렬된다[3].

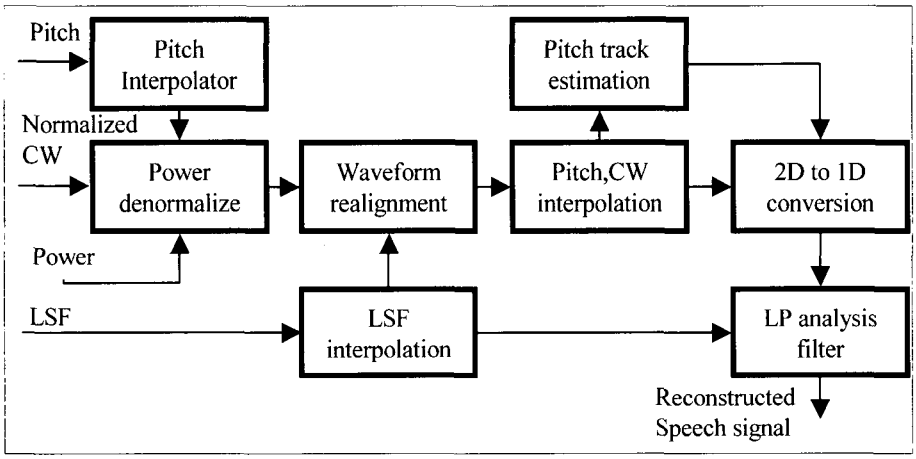
3.3.4. 특성파형 전력 계산 및 정규화

특성 파형의 계수를 특성파형의 전력으로 정규화하고 정규화된 특성 파형 계수를 양자화 함으로서 보다 효과적인 압축이 가능하게 된다.

이상의 과정에서 부호화기는 스펙트럼 정보를 담은 LSF 계수와 잔차 신호의 전력, 피치주기와 전력으로 정규화된 특성파형의 일부를 DTFS 계수 형태로 전송하게 된다.

3.4. 파형보간 복호화기의 구조

복호기는 부호화기의 반대 과정을 거치며 <그림 5>에서와 같이 부호화된 피치 주기를 선형 보간한 정보와 전력을 이용하여 정규화된 특성 파형을 복원(denormalization)한다. 그 다음 양자화로 인해 왜곡된 특성 파형의 피크 성분을 재정렬한다. 구해진 특성 파형 및 피치를 이용하여 복호기 단에서의 특성파형을 선형 보간을 통해 구한다. 특성 파형의 선형 보간은 특성 파형 정렬과 유사한 시나리오를 따른다. 동일 차원의 경우 단순 선형보간, 다른 차원의 경우 절단(truncation) 혹은 영 삽입(zero padding)을 수행한 후 선형 보간을 수행한다.



<그림 5> 파형 보간 복호기 블록 다이어그램

다음은 선형 보간된 피치 값으로부터 식 (5)을 이용하여 위상 궤적(phase track), $\phi(n)$을 구한 후 특성 파형과 위상 궤적을 이용하여 잔차 신호, $r(n)$를 복원한다. 잔차 신호를 복원하는 것은 2차원 신호를 다시 1차원으로 변환하는 과정이 되며 식(11)를 이용하여 수행 한다.

$$\phi(n) \approx \phi(n-1) + \pi\left(\frac{1}{P(n-1)} + \frac{1}{P(n)}\right) \tag{5}$$

$$r(n) = s(n, \phi(n)) = \sum_{k=1}^{[P(n)/2]} [A_k(n)\cos(k\phi(n)) + B_k(n)\sin(k\phi(n))]$$

$$0 \le \phi(\cdot) < 2\pi \tag{6}$$

구해진 잔차 신호와 전송된 LSF 계수를 이용하여 LP 합성 과정을 거치면 복원 음성을 얻게 된다.

3.5. 양자화

3.5.1. LSF 양자화

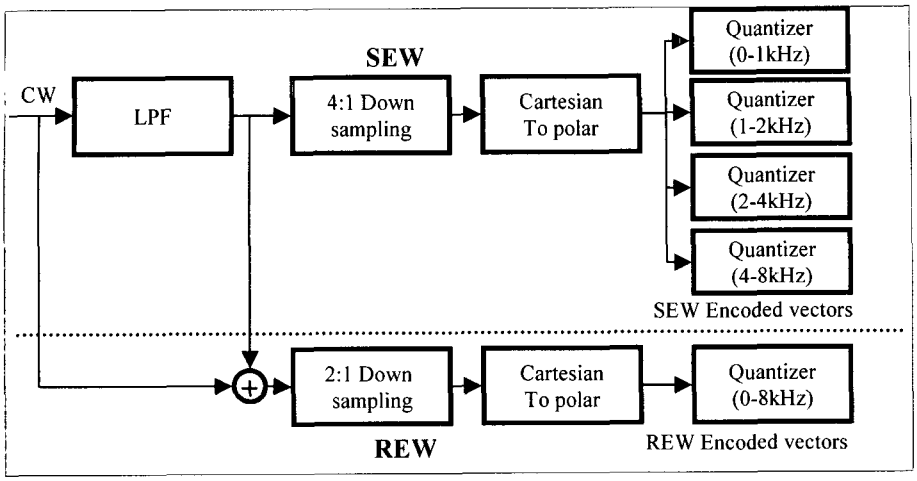
LSF는 분석프레임마다 한번 전송하며 spilt vector 양자화 기법을 이용 LSF 계수를 서브 구간으로 나누고 각각의 서브구간은 독립적으로 양자화 한다. 코드북(codebook)은 주어진 LSF 계수가 스펙트럼 민감도와 포락선에 비례하도록 가중치를 부여하여 MSE(mean square error)를 최소화 하도록 훈련한다[5]. 복원 시 각각의 서브 구간은 순서대로 코드북 검색을 수행하여 서브벡터 순서대로 최적 검색을 수행한다.

3.5.2. 전력(power) 양자화

전력(power)은 SEW(slowly evolving waveform)의 부호화율에 따라 분석프레임별 부호화율이 결정된다. 특성 파형 추출을 8번 하고 매 4번의 특성 파형 중 한번만 보낸다면 전력은 1:4 다운 샘플링하여 보내면 된다. 즉 8개의 서브 프레임 전력 중 2 개만 전송하면 된다. 전력은 최초 로그 도메인으로 변환되고 다운 샘플링으로 인한 스펙트럼 간섭을 막기 위해 저역 필터를 통과시킨 후 다운 샘플링 후 양자화 된다. 복호기에서는 가역 과정을 수행한다.

3.5.3. 특성 파형의 양자화

특성 파형은 양자화하기 전에 먼저 식 (2)를 이용 SEW(slowly evolving waveform)과 REW(rapidly evolving waveform)으로 해체된다. 각각의 파형은 인간의 인지 특성에 따라 독립적으로 양자화한다. SEW의 경우 음성의 유성음 성분이 주를 이루고 피치 주기에 따라 천천히 변화하므로 상대적으로 비트 할당은 많이 하여 양자화 해상도를 높이되 적은 횟수로 전송하는 것이 효과적이고, REW의 경우 무성음 성분 혹은 잡음 성분으로 구성되고 무성음의 경우 시간 영역 해상도가 주파수 영역 해상도보다 중요하다는 연구 결과에 따라[11] 비트 할당은 적게 하되 상대적으로 많은 횟수로 전송하여 음질을 유지하고 압축 효율을 높일 수 있다. 특성 파형 해체(decomposition)에 사용되는 non-causal 저역 필터의 경우 별도의 저역 필터를 사용할 수 있으나 연산 복잡도를 줄이기 위해서 전력 양자화에 사용되었던 동일한 저역 필터를 사용하는 것이 효과적이다. <그림 6>에서 특성파형의 양자화 과정을 도시하였다 복원과정은 역과정을 수행한다.



<그림 6> 특성파형(Characteristics Waveform) 양자화 블록 다이어그램

SEW(slowly evolving waveform)의 경우, 주파수 대역에 따른 인간의 인지특성을 고려하여 대역별로 양자화하는 것이 양자화 효율을 높일 수 있다. REW(rapidly evolving waveform) 양자화의 경우 위상 스펙트럼을 랜덤위상 스펙트럼으로 교체하여도 음성 열화는 적다[3][11]는 결과를 이용하면 위상 스펙트럼을 랜덤 위상 스펙트럼으로 교체함으로서 코딩 효율을 높일 수 있다.

SEW, REW의 코드북 훈련은 LSF 코드북과 유사한 방법으로 수행할 수 있으나 스펙트럼 차원(dimension)이 피치 길이에 비례한다. 따라서 특성파형의 스펙트럼은 가변 차원을 가져야 하고 적절한 VDVQ(variable dimension vector quantizer)를 이용하여 코드북 구성을 하여야 한다.

4. WI 부호기 구현

본 연구에서는 샘플링 주파수 16 KHz, 해상도 16 Bit를 가지는 입력음성에 대한 부호화 및 복호기를 구현하였다.

4.1. 파형 보간 부호기의 구현

부호기의 구현은 3장에서 언급한 순서에 따라 구현 하였다. 16 KHz, 16bits 광대역(wideband) 음성에 대해 분석 프레임은 20msec, 선형 예측 분석 창 함수의 크기는 30msec으로 하였다[1][2]. LPC 차수는 24차로 하고 급격한 스펙트랄 피크 발생 방지를 위하여 0.98829 대역확장(bandwidth expansion)을 수행하였다.

피치주기는 epoch 신호 검출로 일차적으로 얻어서 다시 수작업으로 최적화된

값을 이용하였으며 분석프레임 내 피치 변화를 보상하기 위해 서브 프레임 내의 피치 값은 인접 분석 프레임의 피치 값을 선형보간하여 추정하였다.

특성 파형은 구해진 잔차신호로부터 추출하게 되는데 분석 프레임 당 8회 추출하여 식(4)와 같이 2차원의 DTFS(discrete time Fourier spectrum)로 표현하고 해당 파형은 피치 주기에 따라 변하는 특성 파형의 길이를 피치주기 40, 40~64, 64~80, 80~256의 4개의 구간으로 나누고 각각의 특성 파형이 속하는 구간은 그 구간의 최대값으로 파형의 길이를 정규화하였다. 특성파형 길이의 정규화는 VDVQ (variable dimension vector quantization) 사용을 보상하는 방법으로 이후 코드 벡터 작성을 용이하게 한다. 특성 파형의 추출은 20msec 분석 프레임에서 피치 주기 분석 결과 최소 주기 값이 40 샘플 이상 인 것으로 나타났으므로 주어진 분석 프레임에서 8회 이상 추출하는 것은 이미 추출된 특성파형과 중복되는 결과이므로 8회로 제한하였다.

DTFS 도메인에서 피크 성분이 일치하도록 정렬하고 전력 정규화한 특성파형에 대해 17-tap non-causal FIR 필터를 이용 20Hz 저역필터링을 통해 특성파형을 REW(rapid evolving waveform)과 SEW(slowly evolving waveform)성분으로 해체하였다. 각각의 파형은 무성음 특성과 유성음 특성을 가지며 특성파형을 각 성분으로 해체함으로써 양자화 효율을 높일 수 있다. SEW의 경우는 유성음 특성을 가진다. 따라서 피치 주기의 파형은 시간 영역에서 천천히 변화하므로 시간 영역해상도를 낮추어 부호화하고 신호의 해상도는 높게 하여 부호화 할 수 있다. 반면 REW는 시간영역에서 빠르게 변화하므로 시간영역 해상도를 높이고 신호의 해상도를 낮추는 방식으로 부호화가 가능하게 된다. 따라서 추출된 8개의 특성 파형에서 REW 성분의 경우 400Hz에서 200Hz로 SEW의 경우 cut-off 주파수 50Hz의 17-tap non-causal FIR 필터를 이용 저역 필터링을 하고 400Hz에서 100Hz로 다운 샘플링한 신호만을 전송한다.

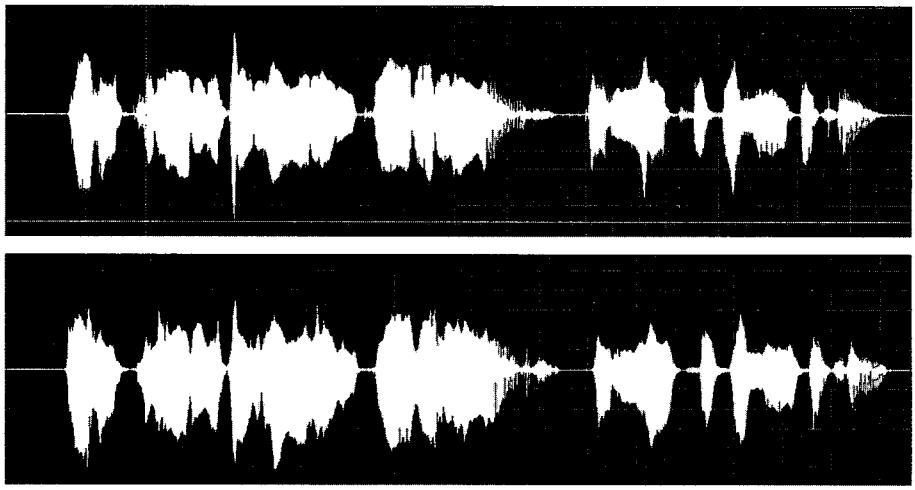
이상의 과정에서 부호화에 사용될 LSF 계수, 잔차 신호의 전력, 피치 주기 및 REW 와 SEW의 DTFS 계수를 구하였다.

4.2. 파형보간 복호기의 구현

복호기는 부호화된 LSF 계수, 잔차 신호의 전력, 피치 주기 및 REW 와 SEW의 DTFS 계수를 이용하여 원 음성을 복원한다. 따라서 부호화기와 동일한 파라미터들(프레임 구조, 버퍼링 등)을 가지며 부호화기의 가역 과정을 수행한다. 단 특성파형을 양자화 하는 경우 양자화 오류로 인해 피크 성분이 부호기에서 정렬된 피크성분이 왜곡되므로 이를 재정렬하고 전송된 피치 값으로부터 위상 궤적을 추정하여 잔차 신호를 복원한다. 복원된 잔차 신호와 전송 이후 복호기 단에서 선형보간 된 LSF 계수를 이용, LP합성과정을 수행하면 최종적으로 음성이 복원되게

된다.

파형 보간 부호기 성능은 전송된 계수를 양자화하지 않고 음성의 복호화를 수행했을 때 복원된 음성의 파형은 <그림 7>과 같이 다소 원음과 파형의 차이는 있었으나 청취했을 때 음질의 차이는 거의 없는 것으로 나타났다. 따라서 파형보간 부호기의 음질열화는 사용된 계수들의 양자화 오류에 전적으로 의존함을 알 수 있다.



<그림 7> 양자화하지 않은 파리미터들로 복원한 음성 파형(위 원음, 아래 복원음성)

4.3. 계수 양자화

파형보간 부호기의 음질열화는 양자화에 의해서만 발생하므로 부호기의 음질은 양자화의 정확도에 따라 결정된다. 먼저 24차로 추출된 LSF 계수는 split vector 양자화 기법을 적용 코드북 훈련을 하였다. 음질향상을 위해 고차의 LPC 파라미터를 사용하는 경우 ,탐색영역 증가로 연산 복잡도는 커지고 요구되는 훈련 데이터는 증가하여 일반적인 벡터양자화 방법으로는 코드북 제작이 용이 하지 않다. 일반적으로 10차 LPC 계수의 경우 100,000 이상의 훈련 데이터가 필요한 것으로 알려져 있다. 따라서 LSF 계수가 몇 개의 그룹으로 스펙트랄 피크의 위치와 대역을 결정하는 것과 나누어진 그룹은 이웃하는 스펙트럼 포락에 영향을 주지 않는 점을 이용[5]하여 LSF 계수를 split 하여 개별적으로 훈련하게 되면 연산복잡도와 훈련 데이타 요구량을 줄일 수 있다. 이때 split 코드는 LSF 계수의 정렬 특성을 위배하지 않아야 안정된 스펙트럼 포락을 얻을 수 있으며 코드북 탐색은 저차의 그룹에서 고차 방향으로 순차적인 탐색을 수행하여야 한다. 코드북 훈련은 스펙트럼 민감도와 포락선에 비례하도록 식 (7)과 같이 LSF 계수에 가중치를 부여하고

MSE(mean square error)를 최소화하는 Generalized Llyod Algorithm을 이용하여 훈련하였다.

$$w_i = |P(f_i)|^r \qquad (7)$$

식 (7)에서 $P(f_i)$는 LPC 전력 스펙트럼이며 가중치 w_i는 주파수 f_i와 실험상수 r에 의해 결정된다. 여기서는 실험치 0.15를 사용하였다.

파형 보간 부호기에서는 피치 정보를 중요시 하므로 피치 주기는 주어진 데이터에 근거하여 최대값을 256 샘플, 최소값을 40 샘플로 설정하여 8 bits으로 표현하고 양자화하지 않았다. 따라서 피치 정보의 손실은 없다.

전력은 SEW 전송과 동일하게 전송되므로 8개의 서브 프레임 중 2개 프레임의 정보만을 전송하므로 1:4 다운샘플링하고 차수 2로 10 bits 벡터 양자화한다.

SEW와 REW는 앞서 언급한 바와 같이 피치 주기에 따라서 DTFS 계수의 차수는 틀려지게 된다. 따라서 적절한 VDVQ(variable dimension vector quantizer)가 필요하다. 각 차수 별 코드북을 만드는 것은 불가능하다. 따라서 여기서는 각각의 가변 차수의 벡터는 아주 큰 벡터에서 일정간격으로 샘플을 취한 것으로 가정하고 주어진 벡터를 고정차수 벡터로 변환하여 코드 북 탐색을 하는 DCVQ (dimension conversion vector quantization)를 수행하였다. 최종적으로 코드북은 하나의 최대 값으로 고정된 차수를 가지게 되는데 본 연구에서는 피치주기의 최대 최소 값 차이를 네 개의 구간으로 나누고, 각각의 구간에 해당하는 벡터를 각 구간의 최대 값으로 맵핑하는 DCVQ를 사용하였으며 따라서 네 개의 고정 차수 코드북을 사용하여 큰 차이의 차수 변경으로 인한 데이터 손실을 방지하였다. 코드북의 훈련은 LSF와 유사하게 split 벡터 양자화 하였는데 4개의 서브대역(1,1,2,4 kHz)으로 나누고 harmonics의 개수가 나누어진 서브대역에 가능한 일치 하도록 각각의 split 차수를 <표 1>와 같이 정하였다.

<표 1> SEW harmonics의 split 차수

	D1	D2	D3	D4
CB1(20)	2	2	5	11
CB2(32)	3	3	7	19
CB3(40)	4	4	8	24
CB4(128)	15	15	30	68

최종적으로 분석 프레임당 각 파라미터의 비트의 할당은 <표 2>와 같다. 따라서 16 비트스트림으로 부호화하게 되면 연속되는 음성의 경우 20 ms, 16 bits의 데이터를 120 bits 비트스트림으로 변화하므로 프레임당 압축률은 42.6배가 되고 6

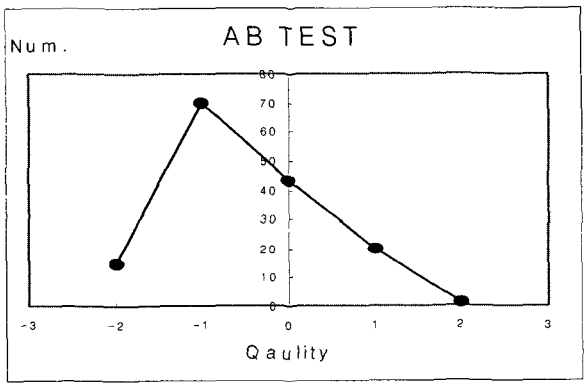
kbps로 표현 할 수 있다.

<표 2> 파형보간 비트 할당

항목	비트 할당
피치 주기	8 bits / frame
LSF	54 bits / frame
전력	10 bits / frame
SEW	36 bits / frame
REW	12 bits / frame

4.4. 성능 평가

부호기의 음질 평가를 위해서 원음과 복원음을 비교하는 AB 실험을 실시하였다. 총 8개 문장에 대해서 20명의 청취자를 대상으로 실험을 실시하였으며 원음과 복원 음을 무작위 순서로 들려주고 어느 쪽의 음이 좋은지 -2에서 2까지 5단계로 평가하도록 하였다. 최종적으로 가장 극단적으로 평가한 두 사람의 데이터를 제외하였을 때 AB test의 결과는 <그림 8>와 같이 나타났다. <그림 8>은 원음에 대한 복원 음의 비교 결과로 가로 축 0 값이 원음과 같다고 평가하는 것이며 양의 값은 우수하다 음의 값은 열화 되었다는 것을 의미한다. 세로축은 평가자 수이다.



<그림 8> 원음(A)과 복원음(B) 비교 실험 결과

평가 결과로 수치상으로 표현하면 원음과 비교하였을 때 복원음이 평균 0.47 정도 열화가 있음을 의미하며, 그 표준편차는 0.82이다.

5. 결 론

본 연구에서는 합성 DB 압축 알고리즘으로 기존의 음성 부호기들을 검토하였고 압축률과 합성 음질에 초점을 맞춘 부호기 성능을 고려하여 파형 보간 방식의 부호기를 구현하였다. 파형 보간 기법의 강점은 첫째 CELP 기반 부호기에서는 실패하기 쉬운 유성음의 주기성을 정확하게 유지하므로 음질 열화가 작다. 둘째 음질의 열화는 파형 보간 기법 자체가 아닌 양자화에 의해서 발생한다는 것이 이미 증명되어 양자화 오차가 없으면 근본적으로 파형보간 기법으로 인한 음질열화는 없다[3][11]. 마지막으로 특성 파형을 REW(rapidly evolving waveform)과 SEW (slowly evolving waveform)으로 분리하여 각각의 파형을 다르게 양자화 함으로서 양자화 효율을 극대화할 수 있다[3].

최종적으로 구현된 파형 보간 부호기는 압축률 42.6배, 원음대비 음질 저하 0.47의 음질 열화를 가진다. 목표한 음질 열화 0.3 이내에 비해 음질 저하가 다소 높은 것은 앞서 언급 한 바와 같이 양자화 오류로 인해 발생한다. 구현된 부호기의 LSF 및 특성 파형의 코드 북 탐색 오류가 음질 열화에 가장 큰 영향을 미치는데 코드 북의 제작은 많은 시간과 훈련 데이타가 필요하고, 또한 데이타에 대한 최적화 작업이 필요한데 이것은 단순히 최적 알고리즘만으로는 해결되기 힘들다. 따라서 자음 모음의 분리 훈련, 음소 간 분리훈련, 복호화시 사용 빈도 수가 낮은 코드워드의 제거 등과 같은 실험에 의한 경험적 지식이 부가되면 음질은 더욱 좋아 질 것으로 예상된다.

구현된 알고리즘을 합성기와 연동하기 위해서는 연속적인 음성렬이 아닌 합성단위별 복호화로 발생 할 수 있는 문제에 대한 고려가 있어야한다. 합성단위 경계처리 및 에러 전파에 대한 처리는 구현된 알고리즘에 추가적인 압축률 저하 및 음질 저하를 초래 할 것이라 예상된다. 그러나 파형보간 알고리즘은 CELP 기반 부 호기와 달리 인접하는 앞선 프레임의 복호화 정보를 이용하지 않는다. 따라서 에러 전파는 CELP 기반 코더만큼 심하지 않으며 합성단위 별 복호 시 경계처리를 위해 합성 유닛의 앞뒤 단 정보만을 이중으로 복호화한다면 현재 압축율을 볼 때 최소 14배의 압축률은 유지 할 수 있으리라 추정되고 이 수치는 현재 DB 압축에 사용되는 알고리즘의 2배의 성능이다. 따라서 본 논문에서는 파형보간 알고리즘의 구현으로 합성 DB 압축에 대한 적합성을 일부 확인하였고 실제 합성 DB 압축과의 연동시스템은 추가적인 연구 과제로 남겨두었다.

참 고 문 헌

[1] C. H. Ritz, I. S. Burnett and J. Lukasiak, "Extending waveform interpolation to wideband speech coding", Proc. of IEEE Workshop on Speech Coding, pp. 32-34, 2002.

[2] C. H. Ritz, I. S. Burnett and J. Lukasiak, "Low bit rate wideband WI speech coding", *Proc. of ICASSP*, Vol. 1, pp. 804-807, 2003.

[3] E. Choy, "*Waveform Interpolation Speech Coder at 4 kb/s*", McGill University, 1998.

[4] G. Kubin, B. S. Atal and W. B. Kleijn, "Performance of noise excitation for unvoiced speech", *Proc. of IEEE Workshop on Speech Coding for Telecommunication*, pp. 35-36, 1993.

[5] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame", *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 1, pp. 3-14, 1993.

[6] M. Ferhaoui, S. Gerven, Lernout *et al.*, "LSP quantization in wideband speech coders", *IEEE workshop on Speech Coding Processing*, pp. 25-27, 1999.

[7] M. Z. Markovic, "Speech Compression - Recent Advances and Standardization", *TELSIKS 2001 5th International Conference on Telecommunications in Modern Satellite*, Cable and Broadcasting Service, Vol. 1, 19-21, pp. 235-244, 2001.

[8] O. Vreeken, N. Pierret, T. David *et al.*, "New techniques for the compression of synthesizer database", *Proc. ISCAS*, Vol. 4, pp. 2641-2644. 1997.

[9] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis, 2nd Ed.*, pp. 3-78, Elsevier, 1998.

[10] W. B. Kleijn, "Encoding speech using prototype waveforms", *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 4, pp. 386-399, 1993.

[11] W. B. Kleijn and J. Haagen, "A general waveform-interpolation structure", *Proc. European Signal Processing Conf.* Edinburg, pp.1665-1668, 1994.

[12] X. Haung, A. Acero and H. HON, *Spoken Language Processing*, pp.337-371, New Jersey: Prentice-Hall PTR, 2001.

접수일자: 2005년 8월 17일
게재결정: 2005년 9월 23일

▶ 양희식(Heesik Yang)
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학원대학교(ICU) 음성/음향 정보 연구실
전화: 042) 866-6296
E-mail: sheik@icu.ac.kr

▶ 한민수(Minsoo Hahn)
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학원대학교(ICU) 음성/음향 정보 연구실
전화: 042) 866-6123
E-mail: mshahn@icu.ac.kr