2003년도 신호처리소사이어티 추계학술대회 논문

# 지능형 고품질 서비스를 위한 오디오 개발

송재종, 이석필, 장세진

전자부품연구원 디지털미디어 연구센터
**Korea Electronics Technology Institute**
**Digital Media Research Center**

# Implementation of The Audio for HiMCS System

Phone: (031) 780-7025
Fax: (031) 780-7060
E-mail: jcsong@keti.re.kr, lspbio@keti.re.kr, sjj@keti.re.kr

## Abstract

본 논문에서는 디지털방송과 인터넷의 융합에 따른 MPEG-2/4/7 방송 및 인터넷 콘텐츠를 비롯한 게임등과 같은 다양한 멀티미디어 서비스를 제공하기 위한 차세대 지능형 고품질 홈 엔터테인먼트 시스템 Platform 개발에서 사용될 MPEG-4 오디오를 개발 한다. 인터넷 상에서의 스트리밍 서비스를 위해서는 저 전송률과 고 품질의 비디오/오디오 알고리즘이 필요하다. 이러한 서비스를 제공하기 위하여 MPEG-4 오디오는 음성에서 고품질의 다중 채널의 오디오까지, 그리고 자연음(Natural Sound)에서 합성음에 이르기까지 다양한 알고리즘을 제공한다. 본 논문에서는 지능형 고품질 미디어 에이전트 시스템에 적합한 MPEG-4 AAC, MPEG-1 Layer-3인 MP3, G.723.1을 구현하고, 이 시스템에 알맞은 7kHz 대역폭을 가지는 광대역(Wideband) 음성신호를 16kbps로 압축하는 음성 압축기를 제안 및 개발한다.

## I. 서론

전 세계적으로 정보통신과 가전분야, 그리고 방송, 통신, 인터넷 분야의 구분이 사라지고 있는 현상이 급속도로 번져가고 있다. 이러한 현상은 다양한 형태의 콘텐츠를 단일 플랫폼에서 처리할 수 있는 시스템의 개발을 필요로 하고 있다. 본 논문에서는 이러한 요구에 따라 방송 및 인터넷상의 다양한 멀티미디어를 처리하고 관리할 수 있는 지능형 고품질 미디어 에이전트 시스템을 개발하고, 이 시스템에 필요한 오디오 및 음성 압축기를 개발 한다. 현재 디지털 방송에서는 AC-3, MPEG-2 AUDIO가 쓰이며, DMB에서는 MPEG-4오디오가 표준으로 채택 되었다. 인터넷상에서 이루어지는 VOD, Streaming과 같은 대부분의 서비스에서 MEPG-4 오디오는 쓰이고 있다. 대부분의 음성 서비스에서 협대역 음성 압축기가 사용되고 있으나, 인터넷(VoIP), 원격회의 등과 같이 고 품질의 음성 서비스를 필요로 하는 분야에서는 광대역 음성 압축기의 필요성이 증대되고 있다. 이러한 요구에 따라 ITU와 3GPP등과 같은 표준화 협회에서는 광대역 음성 압축기의 표준화 작업이 진행 중이다.

이와 같이 다양한 응용분야에서 고 품질의 오디오와 음성 서비스가 요구됨에 따라 본 논문에서는 고품질 오디오 구현과 저 전송률의 고 품질의 음성 통신을 할 수 있는 새로운 광대역 음성 압축기를 제안하고 개발한다.

## II. 지능형 고품질 서비스를 위한 오디오

지능형 고품질 미디어 에이전트 시스템은 다양한 방식의 멀티미디어를 처리할 수 있다. 따라서 각각의 서비스를 위한 오디오 형식이 필요하게 된다. 디지털 방송 수신을 위해서는 AC-3가 필요하게 되고, VOD

서비스에서는 MPEG-4 AAC나 MPEG-1 오디오가 필요하게 된다. 뿐만 아니라 인터넷 화상 서비스나 VoIP서비스를 위해서는 G.723.1이 필요하게 된다. 독자적인 형태의 서비스를 위해 새로운 방식의 오디오/음성 압축기가 필요하다.

### 2.1 MPEG-4 AAC

MEPG-4 AAC는 5개의 객체(AAC Main Object, AAC LC Object, AAC SSR Object, AAC LTP Object, AAC Scalable Object)가 표준화 되었다. 각 객체는 MPEG-2 AAC 프로파일을 기본으로 하여 몇 가지의 툴을 첨가하여 이루어 지기 때문에 MPEG-2와 유사한 기능을 가지게 된다. 각 객체는 약간의 차이를 가지지만 공통된 모듈을 가지게 된다. AAC 객체들이 가지고 있는 공통된 모듈에 대한 기능은 다음과 같다[4].

이득제어(Gain Control), MDCT(Modified Discrete Cosine Transform), TNS(Temporal Noise Shaping) : Intensity Stereo , M/S Stereo , Prediction , Scale Factor, Quantization, Noiseless

### 2.2 MPEG-1 Layer-3

MPEG1 오디오의 경우는 Layer-1,2,3 세 가지 동작 모드를 갖는다. Layer의 수가 올라갈 수록 압축율이 좋지만 복잡성이 커진다. Layer-3는 32 ~ 320 kbps의 Bit rate가 필요하며, bitstream은 프레임(frame) 단위로 구성되어 있으므로 다양한 용도로 조작이 간편하다. 입력 신호는 Subband라 불리는 일정 수의 주파수 밴드로 나뉘어 Coding하는데 여기서 발생하는 양자화 잡음(Quantization Noise)이 Subband의 Masking curve 범위를 벗어나지 않는 범위에서 양자화(Quantize) 한다. 여기서 양자화 잡음의 스펙트럼은 입력 신호의 스펙트럼에 동적으로 적용되게 한다. 각 Subband에 쓰인 Quantizer에 대한 정보는 코드화 된 Subband sample과 같이 전송을 하게 된다. Decoder는 Encoder가 어떻게 이 정보를 탐지했는가를 알지 않더라고 Bit stream을 Decode할 수 있다. 이 때문에 Encoder들은 서로 다른 품질과 복잡성을 가지게 된다[5].

### 2.3 G.723.1

대역폭이 넓은 오디오 신호는 일반적으로 주파수영역에서의 부호화를 이용하는 반면 4kHz의 대역폭을 가지는 음성신호는 사람의 발성 기관을 수학적인 모델링을 바탕으로 한 CELP의 압축 방식을 이용하게 된다. 이는 사람의 음성을 가장 효과적으로 압축할 수 있는 방법으로 여겨지고 있다. G.723.1은 6.3kbps와 5.3kbps를 가지고 있으며, 5.3kbps는 현재 가장 많이 사용되고 있는 ACELP방식을 기반으로 하고 있으며 6.3kbps에 비해 낮은 음질을 제공하지만 좀더 간단한 구조를 가지고 있다는 장점이 있다. 이 부호화기는 30 ms를 프레임 단위로 부호화하게 되므로 일반적인 20ms단위로 부호화하는 음성압축기에 비해 지연이 긴 편이다[6].

## III. 제안하는 오디오 압축기

본 논문에서는 AMR의 12.2kbps~7.40kbps 모드와 G.722.1을 변형하여 새로운 방법으로 광대역 음성신호를 압축하게 된다. 제안하는 방식의 개념도는 그림 1과 같다. 입력신호의 형태는 16Bit PCM이고, 프레임의 길이는 20ms(320 sample)이며, Look-Ahead는 20ms이다.

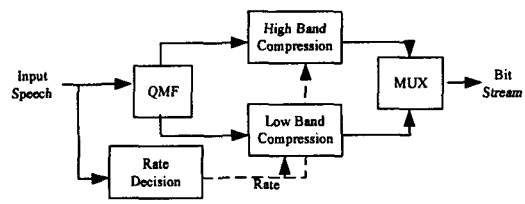


그림1. 제안한 압축기의 개념도

압축기의 전체적인 구조는 입력신호로부터 상위/하위 밴드에 Bit할당을 결정하는 부분, 상위/하위 밴드로 나누는 부분, 상위/하위 밴드를 압축하는 부분, 마지막으로 상위/하위 밴드의 파라미터들을 직렬화하는 부분으로 나누어진다. AMR은 전체 8개 모드로 구성 되고, 4.75kbps~ 12.2kbps까지 지원된다. 입력신호를 차단주파수가 80Hz인 고역 통과 필터를 적용하여 원하지

않는 저주파 성분을 제거하고, 10차 LPC 합성 필터로 음성 스펙트럼의 Envelope을 추정하고, LPC 계수를 LSP로 변환하여 양자화한다. 합성 필터의 여기신호는 피치의 정보를 가지는 적응 코드북과 여기신호인 고정 코드북으로 나누어서 해석한다. 적응 코드북은 다음식을 최소로 하는 k만큼의 지연을 가지는 과거의 여기신호가 된다.

$$e = x[n] - g_p y_k[n] \quad n = 0,1,\cdots,39 \qquad (1)$$

여기서, $g_p$는 적응 코드북의 이득이고, x[n]은 목표 신호이고, $y_k$[n]은 합성된 신호이다.

고정 코드북은 여러 개의 펄스로 이루어진 고정 대수 코드북이며, 각 모드 별로 펄스의 개수가 다르다. 최적의 고정 코드북은 아래 식을 최대로 하는 펄스의 위치와 크기를 선택함으로써 고정 코드북이 정해진다.

$$A_k = \frac{(C_k)^2}{E_{Dk}} = \frac{(\sum_{n=0}^{N-1} x'[n]z[n])^2}{\sum_{n=0}^{N-1} z[n]z[n]} = \frac{(d^t c_k)^2}{c_k^t \Phi c_k} \qquad (2)$$

여기서, x'[n]은 피치의 역할을 제외한 목표 신호이고, z[n]은 고정 코드북에 의해 합성된 신호이다.

G.722.1의 기본 구조는 한 프레임 길이가 20ms로 대역폭이 50Hz ~ 7000Hz이다. Look-Ahead는 20ms를 가지고, 전체 알고리즘 지연은 40ms이며, 전송률은 24kbps와 32kbps를 지원한다[2].

입력신호를 MLT 변환하여 500Hz씩 16개 밴드로 나누어서 주파수 영역에서 압축하게 된다. MLT 변환은 Type IV DCT로서 기본 함수에 대해 Overlap and Add로 완전 복원(Perfect Reconstruction)이 가능하다. 변환된 MLT계수를 16개 밴드로 나누어 RMS을 구한 후, RMS에 대한 Log 양자화 Index를 구하고, 각 밴드별 Index 차이를 Huffman Coding하게 된다. 밴드별 Bit 할당 방법은 밴드별 RMS을 이용해 매우 간단하게 정한다. 각 밴드별로 할당된 Bit수는 직접 전달하지 않고, 각 밴드별 Bit수의 분포에 대한 패턴을 16가지로 나누고, 이중에서 실제로 사용하는 패턴에 대한 Index만을 전달한다. 16가지 패턴 중에 어느 패턴을 사용할지는 모든 경우에 대해 양자화를 실시한 후에 최종적으로 사용할 패턴을 정한다. 16가지 Bit 패턴은 밴드별 RMS만으로 결정되므로 패턴 Index만 보내면 모든 밴드에 대한 Bit수가 전달 된다[2].

### 3.1 상위/하위 밴드 Bit 할당

기존의 압축기는 상위 밴드와 하위 밴드의 비트 할당이 항상 고정된 반면, 제안하는 압축기는 상위 밴드와 하위 밴드의 비트 할당이 입력신호의 특성에 따라 변하게 된다. 입력신호를 Hamming Window를 씌운 다음, 256-Point DFT을 이용하여 파워 스펙트럼을 구하여 상위 8개 블록과 하위 8개 블록으로 전체 16개의 블록으로 나누고, 각 블록별 RMS를 구한다. 구해진 RMS의 블록별 차이에 Weight를 주어 현재 프레임의 상위밴드와 하위밴드의 Bit를 결정하게 된다.

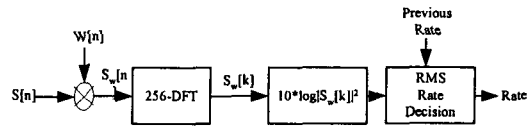


그림2. Rate결정

### 3.2 상위밴드 압축

상위밴드 압축은 입력신호를 MLT을 이용해서 변환하는 대신 Wavelet를 이용하여 변환하게 된다. MLT는 정적 신호에 좋은 결과를 나타내는 반면, Wavelet은 Transient 신호에 더 좋은 특성을 나타낸다. 본 논문에서는 12 tap-Daubechies 직교 필터를 사용하고, 3레벨의 Wavelet Packet를 이용한다. Wavelet Packet를 이용하여 High를 Low/High로 분리 할 때, 2:1 Decimation에 의한 Low/High 순서의 뒤바뀜에 주의하여야 한다. 그림4는 급격히 변화하는 신호에 대하여 Wavelet 계수가 MLT 계수보다 Energy Compaction이 우수함을 보인 것이다. 각 밴드별 Wavelet 계수의 양자화 과정은 G.722.1에서 사용하는 Category방식을 사용한다. 양자화된 RMS를 이용하여 각 밴드에 대한 Category를 정하고, 전체적인 Category분포의 패턴을 총 16가지를 구한 후, 가장 적절한 패턴을 이용해서 양자화를 수행한다.

### 3.3 RMS양자화 방법

제안된 음성 압축기는 입력신호를 하위/상위 밴드로 나누어서 압축하기 때문에 상위밴드의 신호들은 상대적으로 크기가 작게 되고, 적은 Bit가 할당되므로 효율적인 압축방법이 필요하다. 상위밴드 압축은 크게 두 부분으로 나눌 수 있는데, RMS 양자화와 Wavelet계수 양자화 부분이다. 새로운 방법의 양자화 개념도를 그림 5에 도식화하였다. 각 밴드의 RMS중 최대값을 구하여 6-bit Scalar 양자화를 실시하고, 양자화된 최대값으로 각 밴드의 RMS을 정규화 한 후, 정규화 된 값들에 대해 5-bit Vector양자화를 실시한다.

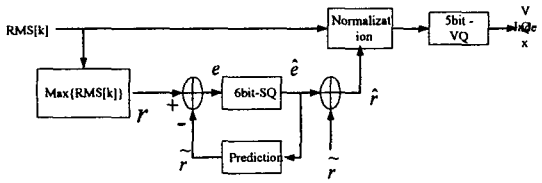


그림3. RMS 양자화 개념도

최대값을 예측하기 위해 4차 MA Filter를 사용하였다. 다음 식은 예측기를 설명한 식이다.

$$\tilde{r}_k = \sum_{n=0}^{4} c[n]\hat{e}[k-n] \qquad (3)$$

여기서, $\tilde{r}$ 는 예측된 최대값을, $\hat{e}$ 은 양자화된 오차값을, c[n]은 예측기의 계수들을 나타낸다. 이렇게 최대값을 직접 양자화하지 않고, 예측오차를 양자화 함으로써 적은 Bit로 효율적인 양자화를 실시할 수 있다. 예측오차에 대한 Scalar 양자화는 다음과 같다.

$$q[n] = 1.13^n \qquad (4)$$

$$q\_bound[n] = 1.13^{n+0.5} \qquad (5)$$

여기서, q[n]은 대표 값을, q_bound[n]은 경계 값을 나타낸다. 양자화된 최대값으로 정규화된 RMS는 5-Bit 7차 Vector 양자화를 실시한다. 이렇게 최대값은 Scalar 양자화를 실시하고, 정규화된 값들은 Vector 양자화를 실시함으로써 양자화기의 SNR 이득이 약 0.6dB 정도 향상된 결과를 얻었다. RMS 양자화를 새롭게 함으로써 양자화기 자체에 대한 이득과 RMS 양자화에 필요한 Bit수가 줄어들어 Wavelet 계수 양자화에 많은 Bit가 할당됨에 따라 전체 음성 압축기의 성능이 향상되었다.

## IV. 성능 평가 및 결론

본 논문에서는 지능형 고품질 서비스를 위한 시스템을 위한 오디오를 구현하고 새로운 방식의 음성 압축기를 제안하였다. 입력신호의 특성에 따라 하위/상위밴드에 가변적으로 Bit를 할당하고, 급격하게 변하는 신호에 대해 Wavelet변환을 적용하고, 새로운 양자화 방법을 제안했다.

| | Good | Fair | Bad |
|---|---|---|---|
| Wavelet | 40% | 50% | 10% |
| MLT | 30% | 50% | 10% |

표1. 제안된 압축기의 음질 평가 결과

음질 성능 측정 결과 G.722 48kbps 음성보다 MLT 방식이나 Wavelet 방식을 사용할 때 모두 제안한 음성 압축기의 성능을 나은 것을 확인하였다.

## 참고문헌

[1] 3GPP TSG, "AMR speech Codec; Transcoding functions", 3G TS 26.090 version 3.1.0, 1999.

[2] ITU-T Rec. G.722.1 "7kHz Audio-Coding at 24 and 32kbps for Hands-Free operation in system with low frame loss", 2000.

[3] ITU-T Rec. G.722 "7kHz Audio-coding within 64kBit/s",1988.

[4] ISO/IEC 14496-3 Part 3 Audio

[5] ISO/IEC 11117-3 Part 3 Audio

[6] ITU-T Rec. G.723.1, Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s, March 1996.