논문 2013-50-1-11

MPEG 통합 음성/오디오 코덱을 위한 오픈 소스 부호화 기술에 관한 연구

(Research on Open Source Encoding Technology for MPEG Unified Speech and Audio Coding)

송 정 욱*, 이 준 일**, 강 홍 구*

(Jeongook Song, Joonil Lee, and Hong-Goo Kang)

요 약

통합 음성/오디오 부호화기 (Unified Speech and Audio Coding, USAC)는 2011년 MPEG에서 FDIS (Final Draft International Standard)를 승인받은 최고 성능의 통합 음성/오디오 부호화기이다. 전통적으로 MPEG에서는 복호화기 기술만 표준화하므로 인코더 기술에 대한 고찰이 쉽지 않을 뿐 아니라, 예제로 공개하는 인코더 (Reference Model, RM)의 경우에도 기본 아이디어만을 포함하고 있기 때문에 이를 사용할 경우 성능 저하가 매우 심각하다. 성능 열화는 매우 심각하다. 이러한 문제를 최소화하기 위해 오픈 소스 기반으로 진행되고 있는 프로젝트 JAME에서는 USAC에 적용된 핵심 인코더 기술의 성능을 최대화 할 수 있는 방법을 제안하고 있다. 본 논문에서는 입력 신호에 따라 두 코더가 선택적으로 동작되게 하는 신호 분류기와 심리 음향 모델을 기반으로 하는 주파수 부호화 기술, 그리고 전이 윈도우 기술 등의 주요 인코더 기술들에 대하여 소개한다. 또한 FDIS를 위한 verification test 결과와 Common Encoder의 성능 평가를 덧붙인다.

Abstract

Unified Speech and Audio Coding (USAC) is the speech/audio codec with the best quality, approved on Final Draft International Standard (FDIS) at MPEG meeting in 2011. Since MPEG conventionally standardizes only the decoder, it is not easy to study on the encoder technologies. Furthermore, Reference Model(RM) shows extremely poor performance. To solve these problems, the open source project(JAME) proposes the methods to make the improved performance of main encoder technologies in USAC. Especially, this paper introduces the encoder modules: the signal classifier for selective operation between two coders, the psychoacoustic model in frequency domain, and window transition technology. Finally, the results of verification test for FDIS and the performance of Common Encoder are appended.

Keywords : Unified Speech and Audio Coding(USAC), USAC Common Encoder (JAME)

Ⅰ. 서 론

통합 음성/오디오 부호화기(USAC, Unified Speech and Audio Coding)는 2011년 7월 FDIS를 승인[1]을 받았으며, 현존하는 오디오와 음성 코덱 가운데 최고의 성능을 가진 통합 표준 코덱이다. 이는 급변하는 멀티미디어 시장의 수요에 대응하기 위해 시작된 표준화 과정으로서 발의된 CfP[2](Call for Proposal)를 시작으로 지난 5년간 지속적인 성능 향상을 통해 얻어진 결과물이다.

종래에 음성 코덱은 오디오 코덱에 비하여 낮은 전송

* 정회원, 연세대학교 전기전자공학과
(School of Electrical and Electronic Engineering, Yonsei University)
** 정회원, LG 전자
(LG electronic Inc.)
접수일자: 2012년6월3일, 수정완료일: 2013년1월3일

률과 짧은 지연 시간의 장점을 활용하여 양방향 통신에 사용되었으며, 오디오 코덱은 폭넓은 대역폭을 제공하며 휴대형 멀티미디어 단말기나 방송통신에 사용되었다. 그러나 최신에 출시되는 다양한 형태의 휴대용 단말기에서는 통신 및 방송에 대한 구분 없이 양질의 콘텐츠를 제공할 필요성이 제기 되었다. 따라서 보편적으로 상용화되어 사용되고 있는 음성 코덱과 오디오 코덱 모두를 하나의 단말기에 포함하여야하는 단점이 있을 뿐 아니라, 많은 경우에는 오디오 코덱을 통해 음성을, 음성 코덱을 통해 오디오 콘텐츠를 주고받으면서 음질 저하의 불편함을 감소해야 했다. 이러한 시장의 움직임을 예측했던 3GPP (Third Generation Partnership project) 에서는 기존 음성코덱인 AMR-WB (Adaptive Multi-Rate Wide-Band)에서 오디오 신호에 대한 음질을 보완하기 위한 TCX (Transform Coded Excitation) 모듈을 추가하여 AMR-WB+ (Adaptive Multi-Rate Wide-Band plus)를 표준화하였다[3]. 또한 MPEG에서도 기존 AAC[4] (Advanced Audio Coding)를 기반으로 하여 저 전송률에서도 음질 저하가 적으며, 낮은 지연 시간을 갖도록 성능을 향상시킨 HE-AAC[5] (High Efficiency Advanced Audio Coding)를 표준화하여 통신 채널을 통해 오디오 컨텐츠를 전달하기 쉽게 하였다. 그러나 여전히 음성 신호에 대해서는 음질이 저조하였기 때문에, 이러한 단점을 보완하고자 새로운 코더인 USAC이 등장하게 되었다.

2007년 CfP이후, 8개의 기관에서 USAC표준화를 위한 후보 시스템을 제안하였으며, 그 중 최고의 성능을 보여주었던 코더를 RM (Reference Model)으로 선정하였다[6]. 이후 수많은 기술이 추가로 제안되었고, 그 중 21개의 CE(Core Experiment)들이 표준안에 채택되었다. 기존의 오디오와 음성 코덱의 저주파 영역 부호화 기술에 해당하는 CE가 8개, 고주파 확장을 위한 SBR (Spectral Band Replication)기술과 스테레오를 위한 CE가 각각 8개, 5개 씩 포함되었다. 이러한 성능 향상에 대한 검증(Veficiation test)을 위한 테스트가 2011년 7월에 진행되었으며, 그 결과 기존의 음성과 오디오 코덱보다 입력 신호에 대하여 일관되고, 뛰어난 성능을 보여주었다[7].

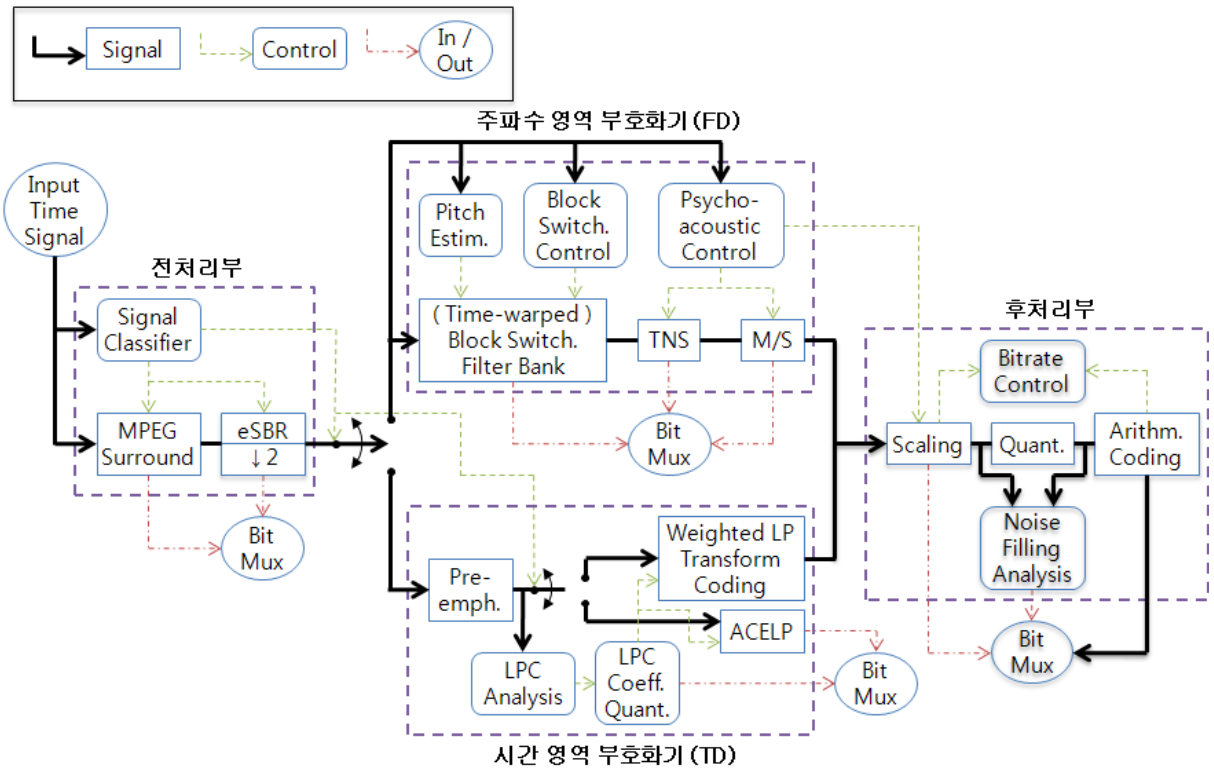

그림 1. USAC 인코더(Encoder) 블록도
Fig. 1. USAC Encoder.

USAC의 표준화는 복호화기에 한정하고 있기 때문에 표준화된 복호화기의 비트열(bit-stream) 규격에 맞는 다양한 형태의 인코더를 설계할 수 있다. 현재 USAC 표준화 과정의 Verification test에서 사용한 RQE (Reference Qualiay Encoder)는 공개가 되지 않고 있으며, 공개된 RM 인코더의 음질은 매우 저조하다. 이러한 문제를 해결하기 위하여 2010년 4월, 92차 회의에서 오픈 소스 기반으로 새롭게 설계된 USAC Common Encoder (JAME)의 구조와 성능 평과 결과가 보고되었다[8~9].

JAME은 본 기관을 중심으로 하여 설계된 소스 코덱이며, 본 논문에서는 USAC 표준화 과정에서 사용된 인코더의 주요 핵심 기술인 신호 분류기(Signal classifier), 윈도우 천이 기술 (Window transition technology), 주파수 영역 부호화 기술, 그리고 무손실 양자화 기술 (Lossless coding)등에 대해 자세히 고찰하고자 한다.

Ⅱ. USAC의 인코더 (Encoder)

그림 1은 USAC의 부호화 과정을 나타내는 블록도이다. 부호화기는 크게 입력 신호가 제일 먼저 들어가는 전처리부와 시간영역 부호화기, 주파수 영역 부호화기, 그리고 후처리부등의 네 가지 부분으로 나누어진다.

MPEG Surround와 eSBR (enhance SBR), 그리고 신호분류기(Signal Classifier)등으로 구성되는 전처리부에서는 입력 신호를 받아 저대역의 모노 신호를 출력한다. MPEG surround에서 스테레오 신호의 위상차 정보를 추출한 후 모노 신호를 출력하고 이 신호가 eSBR의 입력으로 들어간다. eSBR에서는 신호분류기의 결정에 따라 코어 밴드 신호와 고주파 대역의 신호를 나누어 부가 정보(side information)를 추출한 다음 2:1 다운 샘플링 (down-sampling)과정을 거쳐 저주파의 음원을 생성한다. 신호분류기는 입력 신호의 특성을 살펴보고 음성 성분이 많은 신호인 경우에 시간영역 부호화기를 선택하며, 그렇지 않은 경우에는 주파수 영역 부호화기를 선택한다.

음성신호는 신호 분류기의 결정에 따라 시간 영역 부호화기에 의해 주로 부호화 되는데, 주요 프로세싱 과정은 기존의 AMR-WB+와 유사하다. LPC(Linear prediction coding)을 통하여 음성 신호의 포먼트(formant)를 추출하며, 잔여신호의 특성에 따라, ACELP (Algebric Code Excitation Linear Prediction) 방식 또는 wLPT (Weighted Linear Prediction Transform Coding)방식으로 부호화 한다[10]. ACELP로 동작될 때는 기존의 AMR-WB+와 동일한 방식으로 전송 비트율에 따라 코드북 인덱스를 전송하지만, wLPT는 기존의 TCX 에서 DFT(Discrete Fourier Transform)를 사용했던 것과는 달리, MDCT(Modified Discrete Cosine Transform)를 사용하여 압축률을 향상시키고 있다. 또한 주파수 영역 부호화기 방식으로 전환할 때 생기는 문제를 해결하기 위하여 FAC (Forward Aliasing Cancelation) 기술과 FDNS (Frequency Domain Noise Shaping) 기술을 추가하여 보완하였다[11].

주파수 영역 부호화기는 기존의 HE-AAC와 유사한 방식으로 심리 음향 모델(Psychoacoustic model)을 기반으로 마스킹 (masking) 되는 신호에 대하여 양자화비트를 줄이는 방식으로 설계되어있다. 또한 후처리 단에서는 양자화된 스펙트럼을 부호화하는데 CAAC (Context Adaptive Arithmetic Coding) 방식을 사용하여 종래의 Huffman Coding 기술보다 압축률을 높이고 있다[6].

Ⅲ. JAME 인코더 (USAC Encoder)

RM은 표준화 과정에서 디코더의 비트열 생성을 위하여 설계되었기 때문에, 대부분의 핵심 기술들이 포함되지 않았다. 신호 분류기의 경우에 입력 신호의 특성에 상관없이, 반복적으로 시간 영역 부호화기와 주파수 영역 부호화기가 선택되었다. 또한 주파수 영역 부호화기는 심리 음향 모델 없이, 입력신호의 에너지에 비례하여 양자화 비트를 할당하도록 설계되었다. 이 같은 중요 모듈의 부재는 심각한 음질 저하로 이어졌다. RM의 음질 저하를 극복하기 위하여, JAME 인코더에서는 Huawei에서 제안한 신호분류기와, 3GPP에서 배포하고 있는 Enhanced AAC의 심리음향 모델을 재설계하여 포함하고 있다[12, 14].

3.1. 신호 분류기 (Signal Classifier)

신호 분류기는 USAC에서 시간 영역 부호화 방식을 사용할지, 주파수 영역 부호화 방식을 사용할 지를 결

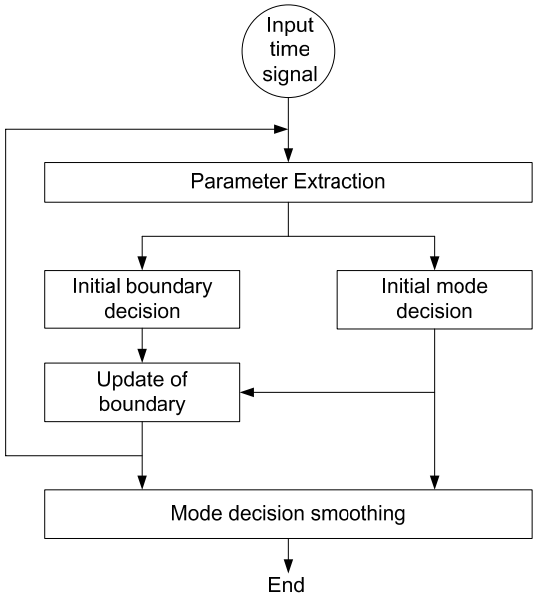

그림 2. USAC 신호분류기 블록도
Fig. 2. Signal classifier.

정하는 중요한 인코더 툴이다. JAME에는 Huawei에서 설계한 신호분류기가 포함되어 있다[12]. 매 프레임마다 부호화 모드가 결정되며, 신호 분류기의 구조는 그림 2와 같다. Tonal 파라미터와 기울기(tilt) 파라미터를 추출한 다음, 음악과 음성의 경계를 찾아 초기 경계를 설정한다. 이것을 바탕으로 초기 모드를 결정하고, 신호 분류 결과에 따라 경계 범위를 갱신한다. 끝으로 에너지, 갱신된 경계범위, 신호 분류 결과를 통해 결정된 모드를 최종적으로 smoothing하는 과정을 거친다. 모드 변환에 따른 경계영역 설정과 갱신, 그리고 모드 smoothing과정은 급격한 모드 전환으로 인한 음질 저하와 비트 낭비를 막기 위해 설계되었다. 본 논문에서는 모드를 결정하기 위한 특성 파라미터의 추출 과정과 이 파라미터들을 사용하여 초기 모드 설정을 하는 과정에 대하여 살펴본다.

3.1.1. Tonal 특성 분석 (Tonal feature analysis)

입력 신호를 FFT(Fast Fourier Transform) 하여 얻은 스펙트럼으로부터 tonal 성분을 추출한다[13]. i번째 프레임에서 전체 대역의 tonal 성분의 수는 $T(i)$로 표시할 수 있으며, 저역 밴드의 tonal 성분의 수는 $T_l(i)$로 표시한다. 또한 긴 구간 tonal 성분 수의 평균은 최근 N_1 프레임 길이의 구간 평균(running average), $\overline{T}(i)$를 말하며 다음과 같이 표현할 수 있다.

$$\overline{T}(i) = \frac{1}{N_1} \sum_{n=i-N_1-1}^{i} T(n) \tag{1}$$

단 구간 tonal 성분 수의 구간 평균은 가장 최근 N_2 개의 프레임의 전체 대역 평균값을 말하며 다음과 같이 표현된다.

$$\overline{T_s}(i) = \frac{1}{N_2} \sum_{n=i-N_2-1}^{i} T(n) \tag{2}$$

저주파에서 tonal 성분의 분포를 알아보기 위한 전체 대역과 저 대역의 비율은 다음과 같이 정의된다.

$$T_r(i) = \frac{\sum_{n=i-N_1-1}^{i} T_l(n)}{\sum_{n=i-N_1-1}^{i} T(n)} \tag{3}$$

3.1.2. 스펙트럼 기울기 특성 분석
(Spectral tilt feature analysis)

i 번째 프레임에 대해 스펙트럼 기울기, $L(i)$는 아래 수식 (4)을 이용하여 구할 수 있으며, 이를 통하여 저주파와 고주파 사이의 기울기를 측정할 수 있다.

$$L(i) = \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} s(n)s(n+1)} \tag{4}$$

이 때 $s(n)$은 입력 신호이고 N는 프레임 길이를 나타낸다.

장 구간의 스펙트럼 기울기에 대한 구간 평균,$\overline{L}(i)$은 아래 수식 (5)을 이용하여 계산할 수 있다.

$$\overline{L}(i) = \frac{1}{N_3} \sum_{n=i-N_3-1}^{i} L(n) \tag{5}$$

장 구간의 스펙트럼 기울기에 대한 평균 제곱근 편차 (mean square deviation), $\tilde{L}(i)$은 아래 수식 (6)을 이용하여 계산한다.

$$\tilde{L}(i) = \frac{1}{N_3} \sum_{n=i-N_3-1}^{i} [L(n) - \overline{L}(n)]^2 \tag{6}$$

단 구간 스펙트럼 기울기에 대한 구간 평균,$\overline{L_s}(i)$은

아래 수식 (7)을 이용하여 계산할 수 있다.

$$\overline{L_s}(i) = \frac{1}{N_4} \sum_{n=i-N_4-1}^{i} L(n) \quad (7)$$

단 구간 스펙트럼 기울기에 대한 평균 제곱근 편차, $\widetilde{L_s}(i)$는 아래 수식 (8)을 이용하여 계산한다.

$$\widetilde{L_s}(i) = \frac{1}{N_4} \sum_{n=i-N_4-1}^{i} [L(n) - \overline{L}(n)]^2 \quad (8)$$

여기서 $N_1 \sim N_4$은 각각 경험적으로 얻은 값이며, 다음 표와 같은 값으로 설정하였다.

표 1. Tonal 파라미터와 스펙트럼 기울기의 길이
Table 1. Lengths of tonal parameter and spectral tilt.

Bit-rate	N_1	N_2	N_3	N_4
12kbps	17.06s	1.706s	13.648s	3.412s
16kbps	12.8s	1.28s	10.24s	2.56s
24kbps	8.53s	0.853s	6.824s	1.706s
48kbps	4.26s	0.426s	3.408s	0.853s

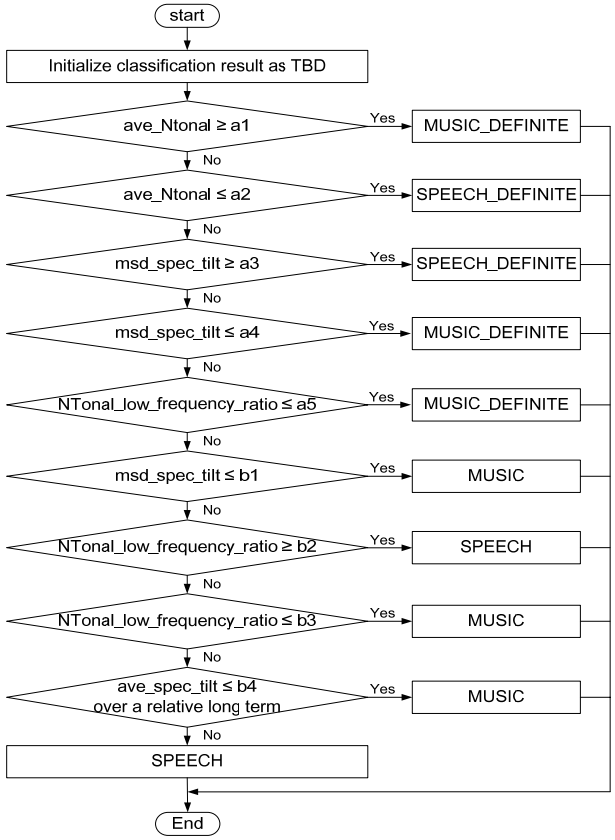

그림 3. 신호분류기의 초기 모드 결정
Fig. 3. Initial mode decision in signal classifier.

3.1.3. 초기 모드 결정

수식 (1), (3), (5), (6)을 통하여 얻어진 스펙트럼 파라미터들은 그림 3과 같이 음성과 오디오를 결정하는 초기 모드에 사용된다. 여기서 MUSIC_DEFINITE와 SPEECH_DEFINITE는 음성과 오디오의 높은 문턱치(thresholds)에 의해 결정된 값이며, MUSIC와 SPEECH는 조금 더 완만한 문턱치에 의해 결정된 값을 나타낸다. 이러한 조건을 만족하기 위하여 동일한 스펙트럼 파라미터의 부등식에서 b_1은 a_4보다 작으며, b_3은 a_5보다 작은 값으로 설정한다. Tonal 성분이 많으며, 스펙트럼 기울기의 평균 제곱근 편차가 작은 신호는 오디오로 분류되며, 그 반대의 경우에는 음성으로 분류된다.

3.1.4. 경계 영역 결정

경계 영역은 크게 음성에서 오디오로 신호가 변하는 경계영역과 오디오에서 음성으로 신호가 변하는 경계영역, 그리고 변화가 없는 경우로 나눌 수 있다. 이 때, 수식 (8)에서 얻은 짧은 구간에서 스펙트럼 기울기의 평균 제곱근 편차와 수식 (2)에서 얻은 Tonal 성분을 추가로 이용하여, 경계 영역을 결정한다.

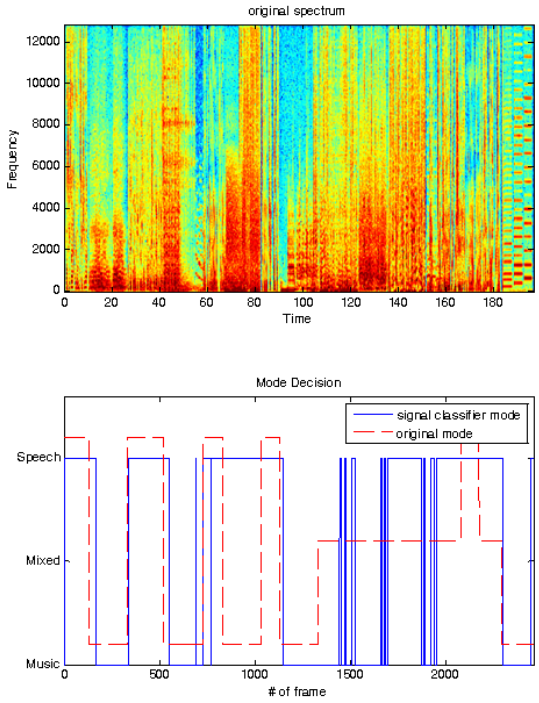

그림 4. 신호 분류기의 모드 결정
Fig. 4. Mode decision of signal classifier.

끝으로 결정된 경계 영역 값과 모드 결정 값을 가지고 과거 이전 $N_5(100)$개의 프레임길이 동안의 음성과 오디오 결정 횟수에 따라 최종 모드를 결정한다. 그림 4는 신호 분류기를 통하여 결정된 모드를 나타낸 그림이다. 파선은 실제 음성, 오디오, 그리고 혼음이 연결된 신호에서 모드를 나타내며, 실선은 신호 분류기를 통하여 결정된 모드이다. 신호 분류기의 smoothing 과정으로 인하여 모드 변경할 때 지연이 있으며, 혼음 신호의 경우에 오디오 성분이 강한 신호는 오디오로 음성 신호 성분이 강한 신호는 음성 신호로 결정하고 있다. 전체적으로 신호분류기의 모드 결정이 실제 모드를 잘 따라가고 있어, 주파수 영역 부호화기와 시간 영역 부호화기의 성능을 극대화 할 수 있다.

3.2. 주파수 영역 부호화기

USAC에 적용된 주파수 영역 부호화기는 심리 음향 모델을 기반으로 한 전형적인 AAC의 구조를 따르고 있다. 입력된 신호를 MDCT영역으로 전환하여, 스펙트럼의 에너지를 바탕으로 마스킹 문턱치 (Masking threshold)를 구한다. 마스킹 문턱치는 저주파로 갈수록 조밀한 형태의 스케일 팩터 밴드 (Scale factor band) 단위로 얻어지며, 스프레딩(Spreading)과 Pre-echo를 방지하는 과정을 거치며, 초기 마스킹 문턱치 값을 얻는다. 이렇게 얻어진 마스킹 문턱치 값은 주어진 비트 전송률에 따라 갱신된다. 한 프레임에서 사용될 수 있는 비트가 한정되어 있기 때문에, 일반적으로 초기 문턱치 값은 상승하게 되는데, 이 때 전 밴드에 동일한 크기의 loudness만큼 더해준다. 즉, 문턱치 값을 조정하는데 있어서 양자화 노이즈를 전 대역에 균일하게 포함시키기 위해 인지적 loudness (perceptual loudness)를 사용한다. 갱신된 문턱치 값, t_r 과 loudness, r과의 관계식은 다음과 같다[14].

$$t_r(n) = (t(n)^{0.25} + r)^4 \tag{9}$$

그림 5는 마스킹 문턱치 값을 조절하는 과정의 예이다. 파선은 원신호의 스펙트럼을 나타내고 있으며, 각 스케일 팩터 밴드 별로 얻어진 초기 문턱치 값이 실선을 나타낸다. 초기 문턱치 값은 수식 (9)에 의해 갱신되며, 1점 쇄선과 같이 조절된다.

마스킹 문턱치 값은 입력신호의 스펙트럼에서 밴드 별로 스케일 팩터(Scale factor)와 전체 이득(Global gain)얻은 데 사용된다. 수식 (10)와 수식 (11)은 스케일 팩터, $c(n)$와 전체 이득, g을 얻는 과정을 설명한다[12].

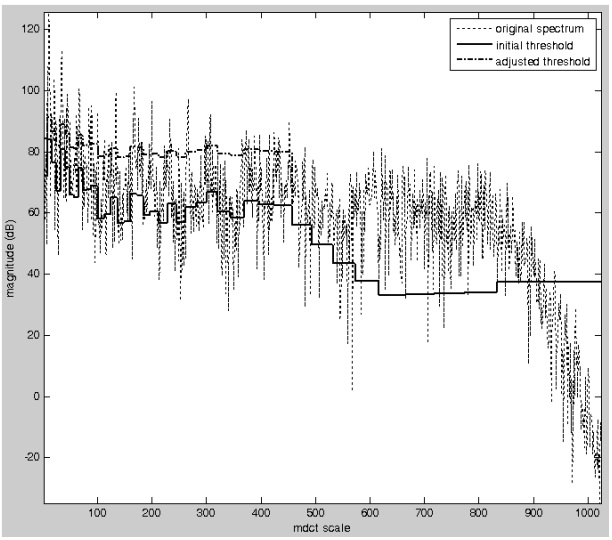

그림 5. 마스킹 문턱치(Masking threshold)의 조절
Fig. 5. Adjusting the masking threshold.

$$g - c(n) = \left\lfloor 8.85\log_{10}\left(\frac{6.75 t_r(n)}{f(n)}\right) \right\rfloor \tag{10}$$

$$f(n) = \sum_{k=k(n)}^{k(n+1)-1} \sqrt{|X(k)|} \tag{11}$$

여기서 $\lfloor x \rfloor$는 산술점 아래 제거 함수를 말하며, $k(n)$은 n번째 스케일 팩터 밴드에 해당하는 스펙트럼의 수를 나타낸다. 전체이득과 스케일 팩터가 정해지면, 스펙트럴 계수(spectral coefficients)는 다음과 같은 수식에 의해 얻어진다.

$$\hat{X}(k) = sgn(X(k))\left[\left(|X(k)|2^{0.25(c(k)-g)}\right)^{0.75} + 0.4054\right] \tag{12}$$

여기서 $sgn(x)$는 부호 함수를 말하며, $[x]$는 Nearest integer 함수를 말한다. 수식 (10)과 수식 (12)에서 알 수 있듯이 스케일 팩터 값과 전체 이득의 차가 크면 클수록 양자화 에러가 크며, 이러한 신호는 마스킹 임계치가 큰 것을 알 수 있다. 결국 심리 음향 모델에 따라 마스킹 문턱치 값을 잘 설계 하는 것이 음질에 커다란 영향을 미친다는 것을 알 수 있다. 그림 6은 12kbps에서 USAC의 RQE와 Common Encoder (JAME)에서 각각 사용된 심리 음향 모델을 이용하여 부호화된 스펙트럼을 나타낸 것이다. Common Encoder

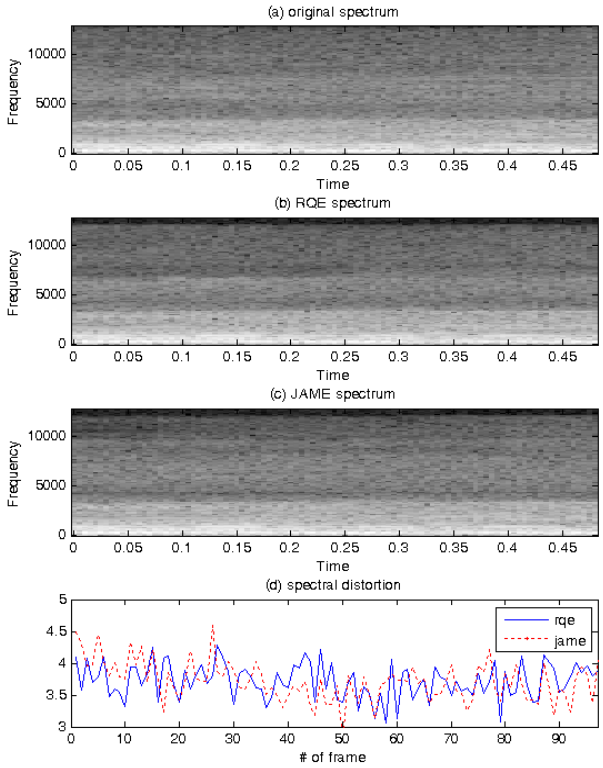

그림 6. (a) 원신호의 스펙트럼, (b) RQE의 주파수 영역 부호화기로 양자화된 스펙트럼, (c) JAME의 코어에서 양자화된 스펙트럼, (d) 스펙트럼 왜곡

Fig. 6. (a) Original spectrum, (b) Spectrum encoded by FD in RQE, (c) Spectrum encoded by FD in JAME, and (d) Spectral distortion

는 3GPP-EAAC의 심리음향 모델을 기반으로 재설계되었다. JAME의 심리음향 모델이 RQE만큼 정교하게 설계되었기 때문에, 그림 6.(d)에서 JAME의 스펙트럼 왜곡정도가 RQE와 차이가 없다. 이 때, 스펙트럼 왜곡정도 (Spectral distance, SD)는 다음과 같은 수식을 통하여 얻는다.

$$SD^2 = \frac{10^2}{2\pi}\int_{-\pi}^{\pi}(\log|H(w)| - \log|\check{H}(w)|)^2 dw \qquad (13)$$

여기서 $|H(w)|$은 원신호의 스펙트럼, $|\check{H}(w)|$은 합성된 신호의 스펙트럼을 나타낸다.

3.3. 시간 영역 부호화기

시간 영역에서 음성 신호를 부호화하는 대표적인 기술은 LPC와 ACELP 그리고 TCX 등이 있다. USAC에는 이와 같은 기존의 기술들이 모두 포함되었으며, TCX의 경우 주파수 영역 부호화기에서 사용되는 MDCT변환 및 무손실 부호화 기술로 CAAC를 이용하여, wLPT라는 기술로 변경되어 적용되었다.

USAC에서 주파수 영역 부호화기와 시간영역 부호화기를 선택적으로 사용하기 위해서 전이 프레임을 설계하는 것은 중요한 문제였다. 우선 주파수 영역 부호화기의 핵심 기술인 MDCT는 현재 프레임의 Aliasing 부분에 다음 프레임의 Aliasing cancelation 부분이 추가되어야 신호의 완벽 복원(perfect reconstruction)이 가능하다. 하지만, 다음 프레임의 모드가 시간 영역 부호화기의 ACELP로 변경된다면, 이러한 Aliasing cancelation 부분을 생성할 수가 없다. 또한 다음 프레임이 시간 영역 부호화기의wLPT로 변경되는 경우에서도, 가중 LPC 합성 필터(weighted LPC synthesis filter)로 인하여 서로 다른 영역에서 연산이 이루어지므로 완벽 복원이 불가능하다.

USAC의 초기 버전에서는 이러한 문제를 해결하기 위하여 주파수 영역 부호화기 다음에 시간영역 부호화기가 올 경우, Aliasing이 존재하는 신호를 버리고, 프레임의 길이를 줄였으며, 다시 주파수 영역 부호화기로 넘어갈 때, 그 길이만큼을 늘려주는 방식을 택하였다[11]. 이러한 방식은 프레임의 길이가 일정하지 않아 불안정한 구조일 뿐 아니라, 버려지는 신호가 존재하여 비효율적이었다.

그림 7은 FAC라는 기술을 이용하여 ACELP로 모드가 변경될 때, 생기는 문제를 해결하고 있다. 그림에서 모드의 경계부분에 화살표 표시되어 있는 영역이 aliasing cancelation 부분이 되며, 이 영역을 주파수 영역 부호화기를 통하여 양자화한 후 추가적으로 전송한다. aliasing 부분은 MDCT와 DCT-IV를 수식 (14)와 같이 정의할 때, 다음 관계식 (15)를 통하여 알 수 있다.

$$\begin{aligned} X_M(k) &= \sum_{n=0}^{2N-1} x(n) w_k\left(n+\frac{N}{2}\right), \\ w_k(n) &= \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right) \\ X_D(k) &= \sum_{n=0}^{N-1} x(n) w_k(n) \end{aligned} \qquad (14)$$

$$\begin{aligned} X_M(k) = &\sum_{n=0}^{N/2-1}\left(-x\left(\frac{3}{2}N-1-n\right)-x\left(\frac{3}{2}N+n\right)\right)w_k(n) \\ &+\sum_{n=N/2}^{N-1}\left(x\left(n-\frac{N}{2}\right)-x\left(\frac{3}{2}N-1-n\right)\right)w_k(n) \\ &=\sum_{n=0}^{N-1} c(n) w_k(n) \end{aligned} \qquad (15)$$

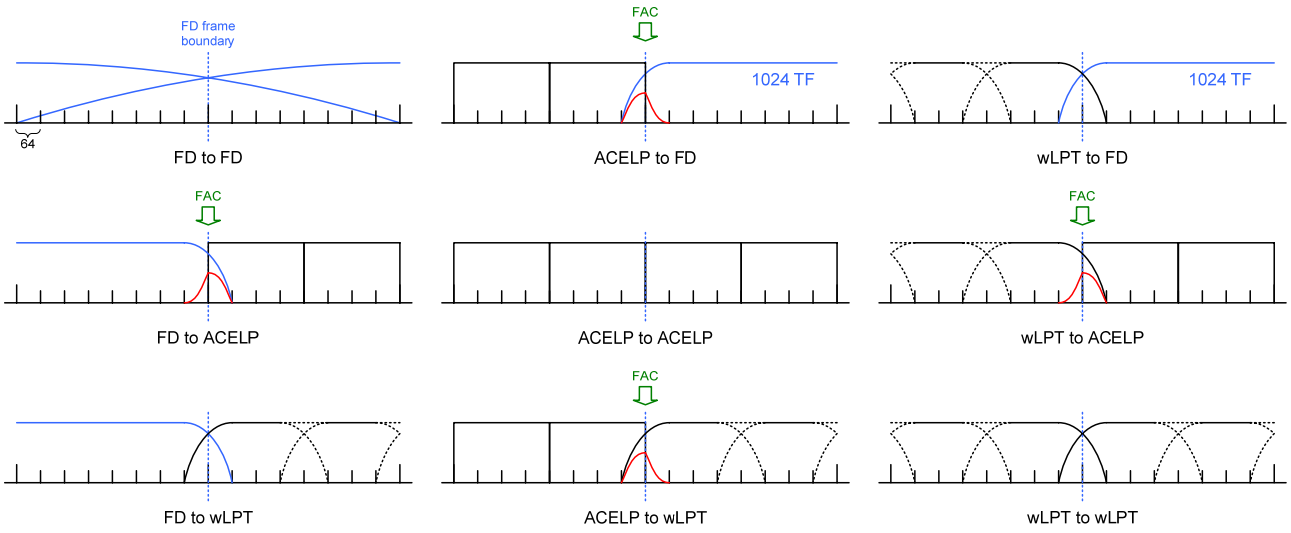

그림 7. USAC에서 Window 전이
Fig. 7. Window transition in USAC.

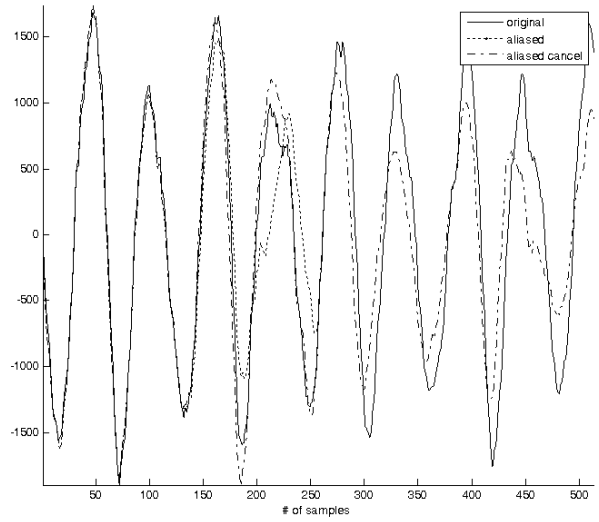

그림 8. FAC를 이용하여 aliasing 제거
Fig. 8. Aliasing cancelation using FAC.

수식 (15)에서 주파수 영역 부호화기 다음에 ACELP가 위치한다고 가정하면, aliasing-cancelation 영역은 $\sum_{n=N/2}^{N-1}(x(n-\frac{N}{2})-x(\frac{3}{2}N-1-n))$와 같은 항이 된다. 즉, aliasing와 alisaing-cancelation 영역은 인접한 두 프레임 사이에서 OLA(overlap and add)의 길이가 N이라고 할 때, 가운데를 중심으로 접혀지는 (folding) 신호 성분이다.

그림 8은 실제 USAC의 전이 프레임에서 FAC를 적용하기 전 aliasing이 생긴 신호(파선)와 FAC를 사용하여 aliasing 부분을 제거한 신호(1점 쇄선)를 나타낸다.

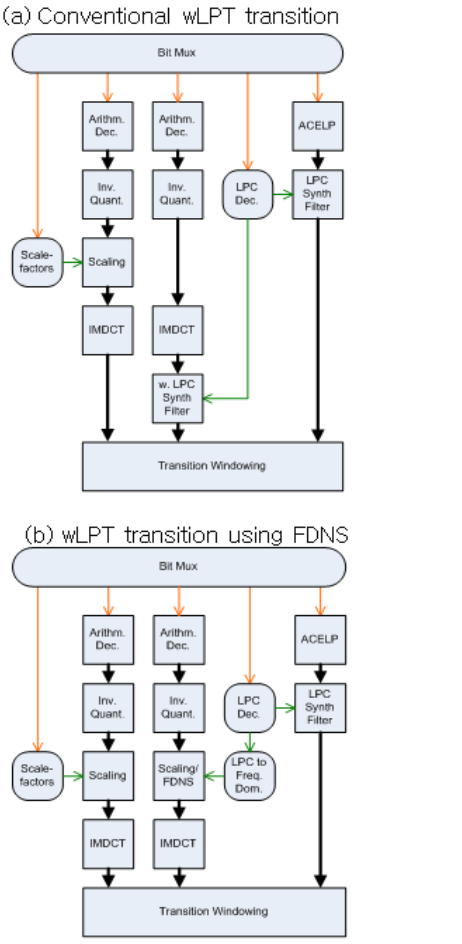

그림 9. (a) 과거 FD와 wLPT 전이 블록도,
(b)FDNS를 이용하여 FD와 wLPT 전이 블록도
Fig. 9. (a) Conventional wLPT transition, and
(b) wLPT transition using FDNS.

파선의 aliasing 부분을 제외한 부분은 1점 쇄선과 일치함을 알 수 있으며, aliasing이 제거 되면서 원 신호(실선)에 가까워지고 있다. 따라서 FAC를 사용하면 전이 프레임에서 생겨나는 왜곡을 제거할 수 있음을 알 수 있다. wLPT와 주파수 영역 부호화기의 전이 프레임에서도 고려할 사항이 있다. 그림 9는 전이 프레임에서의 wLPT의 복호화 과정에 대한 그림으로 FDNS를 적용하지 않은 블록과 적용한 블록을 묘사하고 있다. FDNS를 사용하지 않을 경우 이전 프레임에서 복원된 신호는 역(inverse) MDCT를 거친 시간 영역 신호이며, 현재 프레임에서 가중 LPC 합성 필터를 거쳐 인지적 가중치가 주어진 시간 영역 (perceptually weighted signal domain)의 신호이다. 서로 다른 영역의 신호를 가지고 완벽 복원을 할 수 없기 때문에, 가중 LPC 합성 필터를 주파수 영역으로 변환하여, 역MDCT를 하기 전에 동작하도록 설계해야 한다. 이러한 일련의 과정에서 사용되는 주파수 영역의 가중 LPC 합성 필터를 FDNS라고 한다.

3.4. 무손실 부호화기 (Lossless coding)

ACELP방식외 wLPT와 주파수 영역 부호화 방식에서는 양자화된 스펙트럼을 압축하여 보낸다. 이 때 기존의 AAC에서는 Huffman 코딩을 사용하여, 빈번하게 나타나는 데이터에 대하여 적은 비트를 할당하는 방식으로 코딩 효율을 높였다. USAC에서는 인접한 스펙트럼의 정보를 이용하여 현재 스펙트럼 값을 추정하는 CAAC 코딩 방식을 채택하고 있다.

그림 10은 CAAC의 동작 과정을 설명하고 있다.

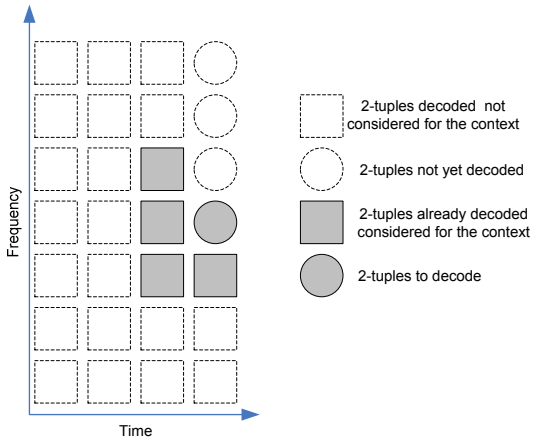

그림 10. CAAC의 동작 과정
Fig. 10. CAAC process.

2-tuples란 인접한 두 개의 스펙트럼 값을 하나로 연결하여 붙인 값을 말한다. 과거 프레임의 인접한 3개의 2-tuples와 현재 프레임에서 인접한 2-tuples를 이용하여 각각의 상위 4비트를 조합하여 코드북 인덱스를 만든다. 코드북 인덱스에 해당하는 값과 현재 스펙트럼 값의 차이가 복호화기에 전송이 된다. 음성과 오디오 신호에 있어서 인접한 스펙트럼의 정보를 이용하는 것은 압축률을 높이는데 있어서 매우 효과적이다.

Ⅳ. USAC의 인코더 성능

1. USAC Verification Test

지난해 7월 FDIS 승인을 위한 USAC의 RQE에 대한 Verification test가 있었다. MUSHRA (MUltiple Stimulus Hidden Reference and Anchor) 테스트[15]를 이용한 주관적 음질 평가는 15개 기관이 참여하여 수행하였다. 모노 저 비트율(Test 1 : 8, 12, 16, 24kbps), 스테레오 저 비트율(Test 2 : 16, 20, 24kbps), 스테레오 고 비트율(Test 3 : 32, 48, 64, 96kbps)로 나누어서 진행된 청취 평가에서 총 24개의 아이템이 사용되었다[16]. 그림 11은 Test 1,2,3에 대한 결과이며, VC(virtual codec)는 USAC 표준화가 시작되면서 만든 기본 요구 사항의 성능을 나타내는 코더이다. USAC기술은 HE-AAC v2와 AMR-WB+, 그리고 VC보다 좋은 음질을 보여주고 있다.

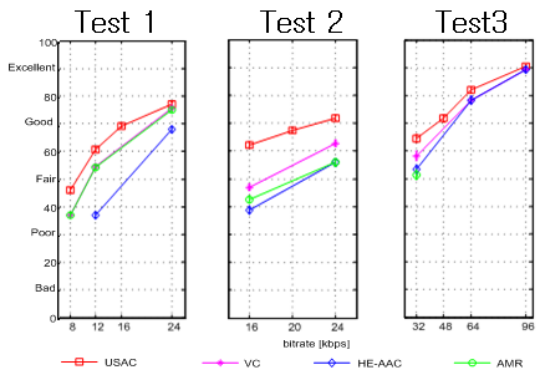

그림 11. USAC verification test 결과
Fig. 11. USAC verification test results.

2. Common Encoder, JAME의 성능

USAC에는 현재 세 가지 인코더가 존재한다. Verification test에 사용되었으나 공개되지 않은 RQE

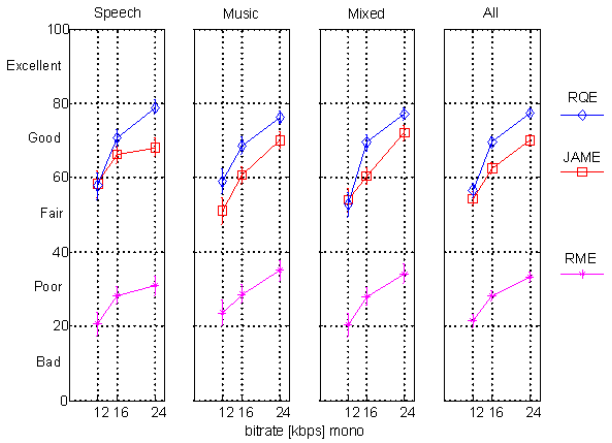

그림 12. JAME의 성능평가 결과
Fig. 12. MUSHRA results in RM, JAME, and RQE.

와 공개된 버전이지만 성능 열화가 심각한 RM, 그리고 RM을 기반으로 성능 개선을 진행 중인 USAC Common Encoder인 JAME이다. 지난 2010년 4월 92번째 MPEG회의에서 open encoder paradigm으로 시작한 JAME project는 LG전자와 ETRI가 협력하여 연세대에서 진행 중이다.

그림 12는 2012년 7월 101번째 MPEG회의에서 보고된 12, 16, 24kbps에서의 성능평가 결과로써 Common Encoder의 성능이 RM보다 훨씬 뛰어날 뿐 아니라 RQE와 유사한 성능을 나타내는 것을 알 수 있다. 12kbps에서 Common Encoder와 RQE는 통계적으로 동일한 성능을 나타내며, 16과 24kbps에서 Common Encoder가 약 7점 가량 낮은 점수를 갖는다.

Common Encoder는 꾸준히 성능개선을 하고 있으며, 기존 SBR에서 enhanced SBR로의 확장과 스테레오 모듈에 대한 확장이 가장 우선적으로 진행될 것이다.

V. 결 론

USAC에서는 복호화기만 표준화하기 때문에 다양한 인코더가 존재할 수 있다. 현재 가장 성능이 뛰어나며, Verification test에 사용된 RQE는 공개되지 않아 현재 최고의 성능을 가진 인코더에 포함된 기술들을 설명하는 것에는 어려움이 있다. 3GPP의 EAAC와 JAME project에서 진행 중인 여러 인코더 기술들을 바탕으로 신호분류기, 심리 음향 모델에 기반한 주파수 부호화 기술, 윈도우 천이 기술, 그리고 무손실 부호화 기술에 대하여 살펴보았다. 또한 현재 RQE의 verification test 결과와 JAME의 성능 평가를 통하여 차세대 음성/오디오 코더 미래를 확인할 수 있었다.

참 고 문 헌

[1] ISO/IEC SC29 WG11 N12231, “ISO/IEC 23003-3/FDIS, Unified Speech and Audio Coding“, 97th MPEG Meeting, July, 2011.
[2] ISO/IEC SC29 WG11 N9519, “Call for Proposals on Unified Speech and Audio Coding”, 82nd MPEG Meeting, October, 2007.
[3] J. M¨akinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, “AMR-WB+: a new audio coding standard for 3RD generation mobile audio services,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’05), vol. 2, pp. II1109 - II1112, March 2005.
[4] K. Brandenburg and M. Bosi, “Overview of MPEG audio: current and future standards for low-bit-rate audio coding,” Journal of the Audio Engineering Society, vol. 45, no. 1-2, pp.4 - 21, 1997.
[5] M. Wolters et al, “A closer look into MPEG-4 High Efficiency AAC,” 115th AES Convention, New York, USA, October 2003
[6] M. Neuendorf, et al., “A novel scheme for low bitrate unified speech and audio coding – MPEG RM0,” in Proceedings of the 126th AES Convention, Munich, Germany, May 2009.
[7] ISO/IEC SC29 WG11 N12232, “USAC Verification Test Report“, 97th MPEG Meeting, July, 2011.
[8] ISO/IEC SC29 WG11 M17571, “Yonsei-LG Contribution to USAC Reference Software”, 92nd MPEG Meeting, Dresden, Germany, April 2010
[9] ISO/IEC SC29 WG11 M23882, “Report on the intermediate verification tests for USAC Common Encoder”, 99th MPEG Meeting, Sanhose, USA, Feb. 2012.
[10] Guillaume Fuchs, et al., “Mdct-based coder for highly adaptive speech and audio coding”, 17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland, August, 2009.
[11] ISO/IEC SC29 WG11 M17020, “Proposal for unification of USAC windowing and frame transitions”, 90th MPEG Meeting, Xian, China, Oct. 2009.

[12] ISO/IEC SC29 WG11 M18470, “A new signal classifier for USAC reference encoder”, 94th MPEG Meeting, Guangzhou, China, Oct. 2010.
[13] ISO/IEC 11172-3:1993, Information technology - Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s, Part 3: Audio.
[14] 3GPP, “General audio codec audio processing functions; Enhanced aacPlus general audio codec; Encoder specification; Advanced Audio Coding (AAC) part”, 2004, 3GPP TS 26.403.
[15] RECOMMENDATION ITU-R BS.1534-1, “Method for the subjective assessment of intermediate quality level of coding systems,” 2001 - 2003.
[16] ISO/IEC SC29 WG11 N12027, “Workplan for Verification Testing of USAC”, 96th MPEG Meeting, Geneva, Switzerland, March, 2011.

저 자 소 개

송 정 욱(정회원)-교신저자
2004년 연세대학교 전기전자 공학과 학사 졸업.
2008년 연세대학교 전기전자 공학과 석사 졸업.
2008년~현재 연세대학교 전기 전자공학과 박사 과정.
<주관심분야 : 디지털 신호처리, 음성 신호처리, 오디오 신호처리>

강 홍 구(정회원)
1989년 연세대학교 전자공학과 학사 졸업.
1991년 연세대학교 전자공학과 석사 졸업.
1995년 연세대학교 전기공학과 박사 졸업.
현재 연세대학교 전기 전자공학과 교수.
<주관심분야 : 디지털 신호처리, 음성 신호처리, 오디오 신호처리>

이 준 일(정회원)
1998년 연세대학교 전자공학과 학사 졸업
2000년 연세대학교 전자공학과 석사 졸업
2000년~현재 LG전자 CTO 부문 재직
<주관심분야 : CPU, DSP, audio 신호 처리>