한국음향학회지 제28권 제7호 pp. 630~639 (2009)

오디오 객체 부호화 표준 - MPEG SAOC

Audio Object Coding Standard Technology - MPEG SAOC

정 양 원*, 오 현 오**
(Yang-Won Jung*, Hyen-O Oh**)

*인텔렉추얼 벤처스 코리아, **LG전자 Digital TV 연구소
(접수일자: 2009년 9월 9일; 채택일자: 2009년 9월 16일)

본 논문에서는 최근 MPEG에서 표준화가 진행되고 있는 오디오 객체 부호화 기술 SAOC (Spatial Audio Object Coding)을 소개한다. SAOC는 이전에 MPEG에서 표준화된 PS (Parametric Stereo), MPEG Surround와 같은 파라메트릭 부호화 기술의 연장선 상에서 특히 오디오 객체 신호를 몇 개의 파라미터를 이용해 부호화함으로써, 사용자에게 음향 장면 구성의 자유도를 제공할 수 있는 객체 기반 서비스에 적합한 기술이다.
핵심용어: 객체 오디오 부호화, 파라메트릭 부호화, 오디오 객체
투고분야: 뉴미디어 분야 (13.3)

This paper introduces MPEG SAOC (Spatial Audio Object Coding) that has been standardized in MPEG audio subgroup. MPEG SAOC is a trendy parametric coding technology conceptually similar to PS (Parametric Stereo) and the MPEG Surround. SAOC especially parameterizes and codes the spatial information for the object signals comprising a downmixed audio scene and thus lets users render one's preferred scene in an interactive manner.
Keywords: Object Audio Coding, Parametric Coding, Audio Object
ASK subject classification: New Media (13.3)

I. 서 론

오디오 부호화 기술의 표준화를 주도하고 있는 MPEG (Moving Picture Experts Group) 내의 audio subgroup의 기술 동향을 살펴보면, 먼저 MP3 로 널리 알려진 MPEG-1 [1]과 이후 등장한 MPEG-2/4 AAC (Advanced Audio Coding) [2][3]를 통하여 심리 음향 모델 (psychoacoustic model)에 근거하여 masking effect를 활용하는 perceptual coding 이 최초의 흐름을 주도 하였음을 알 수 있다.

이후 등장한 SBR (Spectral Band Replication) 기술을 활용한 HE-AAC v1 (High-Efficiency Advanced Audio Coding version 1, 2003년) [4]은 오디오 신호의 대역폭 (bandwidth)를 확장하기 위한 방법으로, 부호화기에서 저대역의 신호로부터 고대역의 신호를 생성하기 위한 parameter를 추출, 전송하여 이용하였고, PS (Parametric Stereo) 기술을 활용한 HE-AAC v2 (High-Efficiency Advanced Audio Coding version 2, 2004년) [5]에서는 역시 부호화기에서 모노 채널로부터 스테레오 채널을 생성하기 위한 parameter를 추출, 전송하는 방법을 제시하였다. 이러한 시도를 통해, 64 kbps의 데이터 양을 갖는 HE-AAC 신호는 128 kbps의 MP3 신호와 필적하는 음질을 제공할 수 있으며, HE AAC v2는 24~32 kbps의 낮은 전송률에서도 양질의 Stereo audio 신호를 제공할 수 있다 [6]. SBR 과 PS 기술의 등장은 low-bitrate에서 고품질의 오디오 신호 전송을 가능하게 하였다는 것 이외에도, 오디오 부호화 기술의 흐름을 perceptual coding 으로부터 parametric coding 으로 전환하는 계기를 제공하게 되었다.

PS 의 성공적인 표준화 이후, MPEG 에서는 하방 호환이 가능한 (backward compatible) 5.1채널 신호의 parametric coding 부호화 방법의 개발에 착수하게 되는데, 이를 통해 탄생한 것이 MPEG Surround로, 모노 신호 혹은 스테레오 신호에 약간의 부가정보 (side information)로써 공간 파라미터(Spatial Parameter)를 함께 전송하여, 수신단에서 다채널 신호의 복호화를 가능하게 한 것이다 [7]. 다시 말하자면, 다운 믹스 신호라 불리우는 모노 혹은 스테레오 신호는 기존의 모노 / 스테레오 수신기 (예를

책임저자: 정 양 원 (ywjung@gmail.com)
140-210 서울시 용산구 한남동 683-73번지 4층 인텔렉추얼 벤처스
(전화: 02-799-3204; 팩스: 02-7799-3250)

들어 MP3 플레이어)에서는 모노 / 스테레오로 재생이 되며, MPEG Surround 기능이 있는 수신기에서는 부가 정보를 활용하여 모노 / 스테레오 신호로부터 5.1 신호를 복호화 하게 된다.

SAC (Spatial Audio Coding)이라는 이름으로 시작된 MPEG Surround 의 표준화는 많은 부분에 있어 C. Faller 의 기념비적인 연구인 BCC (Binaural Cue Coding) [8] 에 근거하고 있다. BCC 는 binaural hearing (양이 청취)의 이론에 근거하여 binaural 효과를 인지하는데 필요한 Binaural Cue 인 ICLD (inter-channel level difference), ICTD (inter-channel time difference), ICC (inter-channel correlation) 를 parameter 로 추출하여 이를 복호화에 적용하는 것을 제안하고 있다. BCC의 application에는 MPEG Surround 와 같은 parametric 다채널 부호화에서부터 flexible rendering, teleconference 등이 제안되었다 [9]. C. Faller 와 F. Baumgarte 가 제안한 flexible rendering 이나 teleconference 등의 응용 분야는 기존의 channel 기반에 머물러 있던 오디오 부호화의 사고를 객체 기반으로 확장시키는 단초를 제공한 것으로, 과거 MPEG-4 표준화의 핵심 개념으로 고려되었으나 여러 물리적, 기술적 제약 때문에 성공하지 못했던 object-oriented coding 개념에 기반한 여러 응용 분야들을 다시 돌아볼 수 있는 계기를 마련하였다.

BCC 라는 기술적 바탕과 PS의 부호화 기법에 기초하여 완성된 MPEG Surround 표준화 이후, 표준 참여자들의 사고는 자연스럽게 object-oriented coding의 패러다임으로 진입하게 되었다 [10].

객체 오디오 부호화 MPEG SAOC가 그 이전의 다른 오디오 부호화 기술과 구별되는 또 다른 개념적 차이는 복호화를 위해서는 다운믹스 신호와 파라미터 비트열 이외에도 복호화기에서 사용자의 입력이 필수적으로 필요하게 되었다는 점이다. 따라서 이를 다루고 처리하는 기술 역시 표준화의 영역 안에 진입하는 계기를 마련하였다.

II. SAOC 표준화 및 APPLICATION

2.1. 표준화 과정

MPEG Surround 의 표준화가 완료되어가는 시점인 2006년 7월, 제77차 MPEG meeting 에서 MPEG Surround 주요 기술권자인 Fraunhofer IIS 등에서 "From Channel-Oriented to Object-Oriented Spatial Audio Coding" [10] 이라는 제목의 기고를 통해 MPEG Surround의 기술을 확장하여 object-oriented spatial audio coding 에 대한 표준화를 고려할 것을 제안하였고, 같은 meeting에서 "Concepts of Object-Oriented Spatial Audio Coding"라는 제목의 MPEG 견의문서를 통해 공식 표준화 의제가 되었다. 그림 1은 MPEG Surround의 채널 기반 부호화로부터 object-oriented 부호화로의 진화를 비교하여 보여주고 있다.

이 객체 기반 부호화 기술은 SAOC (Spatial Audio Object Coding) 의 명칭을 가지게 되었고, 제79차 미팅인 2007년 1월에 CfP (Call for Proposals)이 공개되었고 [11], Fraunhofer IIS/Philips/Coding Technologies 의 공동 proposal과 LG 전자의 proposal이 RM (Reference Model) 선정을 위한 경합을 벌인 끝에 Fraunhofer IIS/Philips/Coding Technologies의 공동 proposal이 2007년 7월 제 81차 미팅에서 MPEG-D SAOC (ISO/IEC 23003-2)의 RM으로 선정되었다 [12].

이후 제84차 미팅 (2008년 4월)에서 CD (Committee Draft) 문서가 공개 되었고, 제86차 미팅 (2008년 10월)에 FCD (Final Committee Draft) 문서가 공개되어 2009년 상반기에 모든 표준화 절차가 완료될 예정이었으나, 표준 기술의 완성도를 높이기 위하여 일정을 조정하여 제89차 미팅 (2009년 7월)에 FCD 문서를 다시 정비 하여 공개하였고 [13], 제91차 미팅 (2010년 1월)에 FDIS (Final Draft of International Standard) 문서를 공개하여 실질적인 표준화 과정을 마무리 지을 예정이다. 이상의 표준화 과정을 통해 성능 향상, 기능 추가 등을 위해, Downmix preprocessor, MBO (Multichannel Background Object),

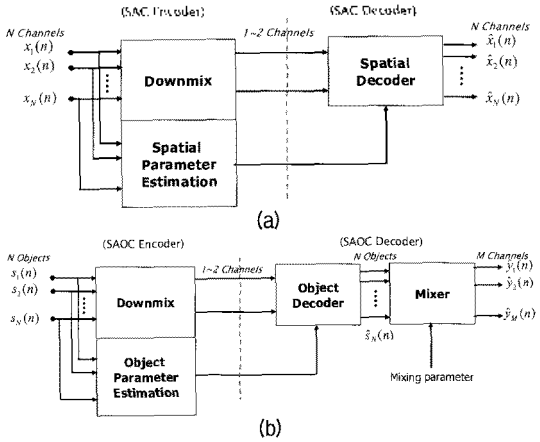

그림 1.(a) 채널 기반 부호화의 블록도 (b) 객체 기반 부호화의 블록도

Figure 1. (a) Block diagram of Spatial Audio Coding (b) Block diagram of Object-Oriented Spatial Audio Coding.

EAO (Enhanced Audio Object), Preset 등의 기술들이 RM에 추가 포함되면서 SAOC encoder 및 decoder architecture 역시 CfP 단계의 RM으로부터 크게 달라지게 되었다.

RM에 선정된 Fraunhofer IIS/Philips/Coding Technologies 이외에도 LG 전자,ETRI, NEC, Panasonic 등의 연구소와 기업에서 현재 SAOC 표준화에 참여하고 있다.

2.2. APPLICATIONS

종래의 오디오 부호화 기술이 주어진 신호의 부호화 / 복호화가 목표였다면, SAOC는 주어진 신호의 부호화 / 복호화 / 재구성이 그 목표가 된다. 이러한 재구성의 필요성과 가능한 응용 분야를 살펴보는 것이 SAOC 기술을 이해하는 데 더 도움이 될 것이라 판단된다.

2.2.1. Interactive Re-mix

SAOC의 가장 큰 응용 분야로 예상되는 것이 바로 Interactive Re-mix이다. 일반적인 음악 컨텐츠는 음악을 구성하는 각각의 악기 (혹은 오브젝트)를 개별적으로 녹음 (각각을 track이라고 한다)한 후, 이것을 믹싱 단계에서 프로듀서의 의도에 따라 적절히 조합하여 만들어진다. SAOC는 이렇게 만들어진 다운믹스에 약간의 부가 정보를 더하여, 사용자가 마치 자신이 프로듀서가 된 것처럼 각각의 오브젝트에 대해 독립적인 제어, 즉 리믹스 (Re-mix) 기능을 제공한다. 예를 들어, 특정 오브젝트의 크기를 줄이거나 키울 수 있고, 특정 사운드 스테이지 안에서 오브젝트의 공간적인 위치를 변경하는 등의 제어가 가능해진다. 극단적으로는 특정 오브젝트의 신호를 제거하는 것도 생각할 수 있다 (보컬 오브젝트를 제거하여 노래방 반주 즉 karaoke에 활용할 수 있다). 또한, MPEG Surround와 마찬가지로 SAOC는 기존 오디오 부호화 포맷과 하방 호환성을 갖고 있기 때문에, SAOC 부가정보가 포함된 다운믹스 오디오 신호는 기존의 일반적인 오디오 재생기를 통한 재생이 가능하다. 아래의 그림 2에서 이러한 interactive re-mix 의 실시 예를 도시하였다.

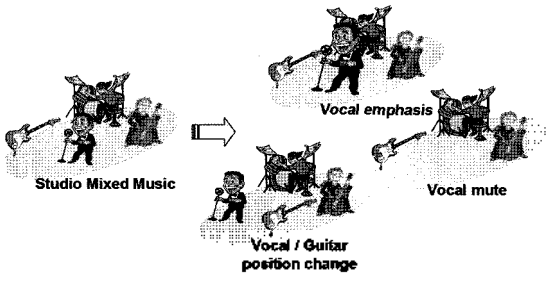

그림 2. Interactive Re-mix.
Figure 2. Examples of Interactive Re-Mix.

2.2.2. Flexible rendering in broadcasting

Interactive Re-mix와 같은 개념을 방송에 적용하는 것이 가능하다. 종래의 단방향성을 가지는 방송에 시청자의 자유도를 제공하려는 시도들이 있어왔는데 [14], 예를 들어 어학 학습용 방송에서 한국어와 외국어의 음성을 선택적으로 청취하거나, 스포츠 중계에서 관중의 소리와 캐스터의 음성 크기의 비율을 취향에 맞게 조절하는 등의 응용 분야가 제시되었었다. 또한, 영화/드라마 등에서 효과음 등 배경 신호의 과도한 음량으로 배우의 대사가 묻히는 경우가 많이 있는데, 이때에도 시청자에게 sound scene 구성의 자유를 제공함으로써 이러한 문제를 해결할 수 있다. 특히 방송의 경우 기존의 object-oriented 방식과 같이 각 트랙에 대해 독립적인 오디오 부호화를 통해 전송할 경우 필요한 대역폭이 전송되는 트랙 수에 비례하여 증가하기 때문에 상업적으로 적합하지 않다. SAOC를 이용하면 증가되는 대역폭이 기존 오디오 신호 대역의 100~150% 정도면 가능하기 때문에 전송 효율면에서 큰 이득이 있으며, 하방 호환성을 보장하기 때문에, 기존 방송 인프라의 부가 서비스로 도입이 용이하며, 차세대 방송 서비스로의 다양한 활용을 기대할 수 있다.

2.2.3. Immersive Teleconference

Interactive Re-mix 와 함께 SAOC 의 주요 응용 분야로 예상되는 곳이 바로 원격 회의 (teleconference)이다. 보통 모노 (혹은 최대 스테레오)로 전송되는 원격 회의 환경의 경우, 원격의 참가자들의 음성이 모두 동일한 물리적 위치 (모노 스피커의 위치)에서 출력되기 때문에, 현재 발화하는 참가자를 식별하기에 어렵고, 화자간 중첩에 의해 대화 내용의 이해에 어려이 있다. 이러한 문제를 극복하고 원격 회의에서 실제감, 현장감을 높이기 위한 연구가 활발히 진행되고 있는데, SAOC는 이와 같은 Immersive teleconference 환경에 활용하기 좋은 부호화 기술이다. SAOC를 이용할 경우, 종래의 모노 전송 채널을 이용하더라도 각각의 원격 참가자의 음성을 가상의 회의실 공간상의 독립된 물리적 위치에서 출력하는 것이 가능하고, 이를 통하여 음성 명료도를 증가시키고 발화하는 화자를 명확하게 식별하는 것이 가능하게 된다. 특히, 화상 회의 (video conferencing)의 경우, 화면상의 원격 참가자의 위치와 음성 재생 위치를 공간적으로 일치시킴으로써 실제감을 높이는데 활용될 수 있다. 이와 같은 상황을 그림 3에 나타내었다.

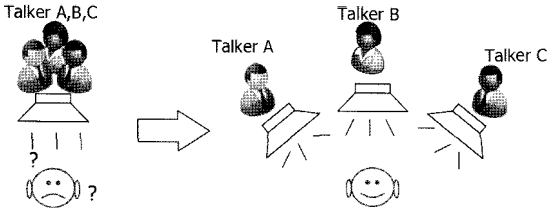

그림 3. Immersive Teleconference.
Figure 3. Example of Immersive Teleconference.

2.2.4. Gaming/Rich media

Gaming 혹은 Rich Media의 경우도 앞선 사례와 마찬가지로 SAOC의 활용 가능성이 높은 응용 분야이다. 게임의 가상 공간속 player간의 대화나 효과을 각각의 오디오 객체로 표현하여 SAOC로 전송하게 되면, 전송 효율에서 매우 유리한 장점을 가진다.

III. SAOC TECHNOLOGY

본 장에서는 MPEG SAOC 부호화, 복호화기, 그리고 주요 기술에 대해 설명한다.

3.1. ENCODER

SAOC 인코더의 구조는 아래와 같다.

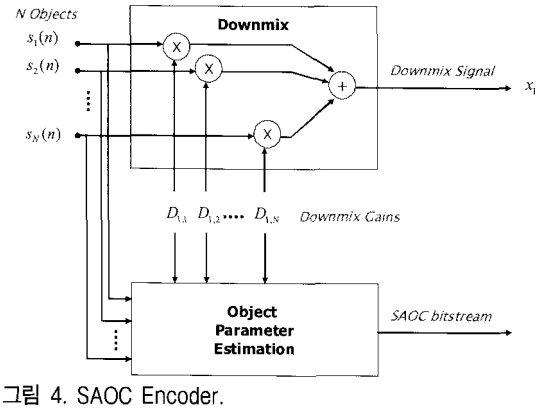

그림 4. SAOC Encoder.
Figure 4. SAOC Encoder.

오브젝트 신호 $s_i(n)$는 아래와 같이 다운믹스 게인에 의해 다운믹스 신호 $x_j(n)$를 생성하게 된다

$$x_j(n)=\sum_{i=1}^{N} D_{j,i}s_i(n) \tag{1}$$

$$\mathbf{X}=\mathbf{DS}=\begin{pmatrix} d_{11} & \cdots & d_{1N} \\ d_{21} & \cdots & d_{2N} \end{pmatrix}\begin{pmatrix} s_1 \\ \vdots \\ s_N \end{pmatrix}=\begin{pmatrix} x_L \\ x_R \end{pmatrix} \tag{2}$$

여기서 $D_{j,i}$는 i번째 오브젝트 신호가 j번째 다운믹스 채널에 포함되는 정도를 나타내는 다운믹스 게인이며, **D**는 이러한 다운믹스 게인들로 구성된 다운믹스 행렬이다. 그림에서 볼 수 있는 것처럼, SAOC 인코더의 입력 신호는 오브젝트 신호 $s_i(n)$와 다운믹스 게인 $D_{j,i}$이고, 출력 신호는 다운믹스 신호인 $x_j(n)$와 추출된 오브젝트 파라미터로 구성되는 SAOC bitstream이다. 실제 오브젝트 파라미터는 MPEG Surround와 동일하게 최대 28개의 파라미터 밴드라는 주파수 분해된 단위로 생성되며, 양자화된 오브젝트 파라미터들은 역시 MPEG Surround에서 사용되던 파라미터 부호화 tool들을 이용하여 SAOC 비트열로 표현된다.

3.2. OBJECT PARAMETERS

SAOC에서는 다운믹스 신호에 포함되는 오브젝트 신호에 대해 다음과 같은 파라미터를 추출하고 있다. 먼저 오브젝트의 레벨값에 대응하는 OLD (Object Level Difference)와 OLD 생성에 사용되는 정규화 값인 NRG (absolute object eNeRGy), 그리고 각 오브젝트 간의 상관 관계에 관한 값인 IOC (Inter-Object Cross-correlation)가 그것이다 이상의 오브젝트 파라미터는 MPEG Surround와 동일한 domain인 Hybrid QMF를 이용하여 71밴드로 오디오 신호를 분해한 후, 4개에서 28개의 파리미터 밴드로 다시 조합하여 각 밴드에 해당하는 파라미터를 추출하게 된다 [7].

각각의 오브젝트 파리미터의 추출은 다음과 같은 식에 의거하여 이루어진다.

$$OLD_i(pb)=10\log10\left(\frac{\sum_{n}\sum_{m\in pb}s_i^{n,m}s_i^{n,m*}}{NRG(pb)}\right) \tag{3}$$

$$NRG(pb)=\max_{k}\left(\sum_{n}\sum_{m\in pb}s_k^{n,m}s_k^{n,m*}\right) \tag{4}$$

$$IOC_{ij}(pb)=\mathrm{Re}\left\{\frac{\sum_{n}\sum_{m\in pb}s_i^{n,m}s_j^{n,m*}}{\sqrt{\sum_{n}\sum_{m\in pb}s_i^{n,m}s_i^{n,m*}\sum_{n}\sum_{m\in pb}s_j^{n,m}s_j^{n,m*}}}\right\} \tag{5}$$

여기서 pb는 파라미터 밴드를 의미한다.

또한, 다운믹스 신호 생성에 사용된 다운믹스 게인들도 파라미터화되어 비트스트림에 포함되는데, 특정 오브젝트가 다운믹스 신호에 포함된 정도를 나타내는 DMG (DownMix Gains) 값과, 다운믹스 신호가 스테레오 일 경우, 오브젝트의 패닝 정보를 나타내는 DCLD (Downmix

Channel Level Difference) 값을 추출하여 전송하게 된다. DMG와 DCLD는 파라미터 밴드 단위로 추출되는 OLD 등과는 달리, 다운믹스 채널에 대해 전대역에 걸쳐 하나의 값으로 추출되는 특징을 가지고 있다.

$$DMG_i = 10\log_{10}(D_{1,i}^2 + D_{2,i}^2 + \varepsilon) \quad (6)$$

$$DCLD_i = 20\log_{10}\left(\frac{D_{1,i}}{D_{2,i}}\right) \quad (7)$$

이와 같이 추출된 각각의 파라미터들은 각 파라미터가 가지는 특성과 범위에 맞게 적합한 양자화 과정을 거치게 되고, 양자화된 파라미터들은 다양한 lossless 부호화 기법들을 이용하여 효과적인 압축이 수행되어 최종적인 비트열로 생성된다. 이때 lossless 부호화 기법으로 각 파라미터의 시간, 주파수 계열간의 상관성을 이용한 시간 및 주파수 차분 부호화, 파라미터의 대표값을 기준으로 차분 부호화하는 파일럿 부호화 (pilot-based coding) 등이 일차 적용되고, 이와 같이 시간, 주파수 중복성이 제거된 파라미터 인덱스들에 대해 최종적으로 Huffman 부호화를 수행함으로써, 효과적으로 압축된 비트열을 얻을 수 있다. 이와 같은 일련의 lossless 부호화는 MPEG Surround에서 사용되던 tool들을 그대로 수용하고 있다. 이때 SAOC 비트열은 오브젝트 당 약 3 kbps 정도로 비트율을 가진다.

3.3. DECODER

3.3.1. Operation Mode

표준화가 시작되는 시점에서 SAOC 는 독립된 오디오 부호화 / 복호화기가 아닌, MPEG Surround와 결합되어 사용되는 형태로 제안 되었었다 [11]. 이는, MPEG Surround라는 parametric하게 멀티 채널을 재생하는 검증된 툴이 있으므로, 표준화 작업에 대한 노력의 중복 투자를 막기 위해, 다채널 재생 전단계, 즉 출력 신호에 대한 사용자의 interaction은 SAOC에서 처리를 하여 주고, 전송된 SAOC bitstream을 MPEG Surround에서 재생할 수 있는 형태로 transcoding한 후, 최종 신호 재생은 MPEG Surround에서 담당하는 구조로 제약 사항을 만든 것이다. 그러나, 표준화 과정에서 MPEG Surround만을 rendering engine으로 이용할 경우, 오브젝트의 panning에 제약 사항이 발생하는 것을 발견하게 되었으며, 이를 해결하기 위해 새로운 툴로써 downmix processor가 도입되었다 [15]. Downmix processor는 stereo 신호까지의 rendering engine으로 동작할 수 있다.

Downmix processor와 MPEG Surround 두 개의 rendering engine을 가지게 된 SAOC는 application scenario에 따라 SAOC 자체가 최종 오디오 신호를 출력하는 SAOC Decoder Mode와, MPEG Surround가 최종 오디오 신호를 출력하는 SAOC Transcoder Mode를 상황에 맞게 선택적으로 이용하도록 표준이 제정되었다. 이러한 Operation Mode는 출력 채널의 수에 따라 아래 표 1과 같이 결정된다.

표 1. SAOC Operation Mode.
Table 1. SAOC Operation Mode.

SAOC module mode	Output signal config.	# of output channels	SAOC module output	MPS decoder required
Decoder	Mono / Stereo / Binaural	1 or 2	PCM output	No
Trans coder	Multi-channel	> 2	MPS bitstream, Downmix signal	Yes

표 1에서 볼 수 있는 것처럼, 사용자가 의도하는 출력 신호가 모노 혹은 스테레오 신호일 경우, SAOC는 Decoder Mode로 동작하며, SAOC Decoder의 출력은 사용자가 의도하는 형태로 rendering이 된 PCM 신호가 된다. 모노 혹은 스테레오 출력인 경우에 MPEG Surround decoder를 이용하는 것은 연산량 및 성능 측면에서 불필요하게 과도한 요구사항이기 때문에 decoder mode로 동작한다고 생각할 수 있다.

반면, 사용자가 의도하는 출력 신호가 멀티채널 신호일 경우, SAOC는 Transcoder Mode로 동작하며, SAOC Transcoder의 출력은 SAOC에 의해 변경된 다운 믹스 신호와 MPEG Surround bitstream이 된다. 최종적인 멀티채널 PCM 신호를 재생하기 위해서는 MPEG Surround decoder가 필수적으로 필요하게 된다. 하나의 시스템 안에서 SAOC transcoder와 MPEG Surround가 구현되는 경우는 굳이 MPEG Surround를 생성하지 않고, 양자화되지 않은 파라미터 레벨에서 SAOC-MPEG Surround 연결이 가능하다.

3.3.2. SAOC DECODER MODE

SAOC가 Decoder mode로 동작하는 경우의 구조는 그림 5와 같다.

SAOC Decoder의 입력은 그림 5에서 볼 수 있는 것처럼, 인코더에서 생성된 다운믹스 신호, SAOC bitstream과, 디코더에서 사용자 입력으로 들어오는 Rendering Matrix

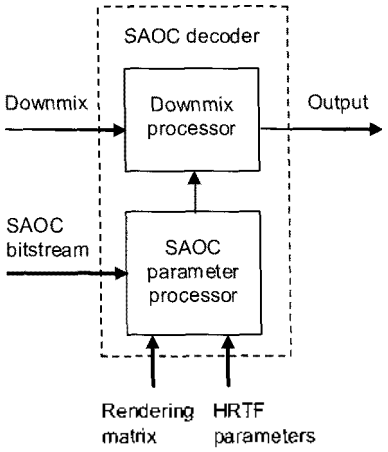

그림 5. SAOC Decoder.
Figure 5. SAOC Decoder.

와 HRTF 파라미터가 있다. 출력은 SAOC로 처리된 PCM 신호가 된다.

서론에서 언급한 것과 같이 디코더의 입력으로 Rendering Matrix (사용자가 원하는 audio scene)을 갖는다는 점은 SAOC가 여타의 오디오 부호화 방법과 다른 차별점이라 할 수 있다. Rendering Matrix는 각 오디오 오브젝트를 원하는 출력 채널에 매핑시키는 것에 관한 것으로, 다음과 같은 수식으로 정의될 수 있다.

$$\mathbf{A}=\begin{pmatrix} a_{1,Lf} & \cdots & a_{N,Lf} \\ a_{1,Rf} & \cdots & a_{N,Rf} \\ a_{1,Rf} & \cdots & a_{N,Rf} \\ a_{1,Lfe} & \cdots & a_{N,Lfe} \\ a_{1,Ls} & \cdots & a_{N,Ls} \\ a_{1,Rs} & \cdots & a_{N,Rs} \end{pmatrix} \tag{8}$$

여기서 $a_{i,ch}$ 는 i번째 오브젝트를 ch (Lf, Rf, … Rs)번째 출력 채널에 할당하는 게인을 나타내는 값이다.

다운믹스 프로세서의 내부 구조는 다음과 같다.

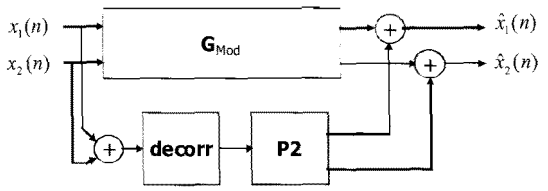

그림 6. Downmix Processor.
Figure 6. Downmix Processor.

다운믹스 프로세서의 입력은 전송된 스테레오 다운믹스 신호 $x_1(n)$, $x_2(n)$ 이고, 출력은 전처리된 신호인 $\hat{x}_1(n)$, $\hat{x}_2(n)$ 가 된다. 그림 6의 과정을 행렬식으로 표현한 것이 아래의 식 (9) 이다.

$$\hat{\mathbf{X}}=\mathbf{G}_{Mod}\mathbf{X}+\mathbf{P}_2\mathbf{X}_d \tag{9}$$

여기서, $\mathbf{G}_{Mod}$ 는 2x2 행렬이고, $\mathbf{X}_d$ 는 디코릴레이터를 통과한 모노 신호, $\mathbf{P}_2$ 는 $\mathbf{X}_d$ 를 믹싱하기 위한 2x1 행렬이다. 여기서 사용되는 디코릴레이터는 MPEG Surround와 동일한 것이 사용된다.

$\mathbf{G}_{Mod}$ 가 산출되는 과정을 개략적으로 살펴보면 다음과 같다. 먼저 모든 오브젝트의 원 신호를 갖고 있을 경우 얻을 수 있는 출력을 $\mathbf{A}_3\mathbf{S}$ (여기서 $\mathbf{S}$ 는 s_i 로 구성된 행렬)라 하면, 이것은 다운믹스 신호인 $\mathbf{X}$ 에 적절한 연산 ($\mathbf{C}_3$ 행렬)을 적용하여 구한 출력과 유사하다고 가정할 수 있다.

$$\mathbf{A}_3\mathbf{S}\approx\mathbf{C}_3\mathbf{X} \tag{10}$$

여기서, $\mathbf{A}_3$ 행렬은 이후 식 유도를 위해 식 (8)의 행렬의 6개의 출력을 3개의 출력, 즉 front left, front right, front center로 할당한 행렬이다.

오브젝트 신호의 공분산행렬은 다음과 같이 정의된다.

$$\mathbf{SS}^*=\mathbf{E} \tag{11}$$

여기서, 식 (2)를 식 (10)에 대입하고, 양변에 $\mathbf{S}^*$ 를 곱하면 아래와 같은 식을 얻게 된다.

$$\mathbf{A}_3\mathbf{E}=\mathbf{C}_3\mathbf{DE} \tag{12}$$

양변에 다시 $\mathbf{D}^*$ 를 곱하고, $(\mathbf{DED}^*)^{-1}$ 를 다시 곱하면 아래의 식을 얻게 된다.

$$\mathbf{C}_3=\mathbf{A}_3\mathbf{ED}^*(\mathbf{DED}^*)^{-1} \tag{13}$$

이를 통하여 $\mathbf{G}$ 는 다음과 같이 얻어진다.

$$\mathbf{G}=\mathbf{D}_{TTT}\mathbf{C}_3 \tag{14}$$

여기서,

$$\mathbf{D}_{TTT}=\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \tag{15}$$

이다.

결국, **G**를 구하기 위해 필요한 것은 $\mathbf{A}_3$, **E**, **D** 행렬인데, $\mathbf{A}_3$는 식 (8)의 형태로 사용자 입력값이 주어지고, **E**는 OLD와 IOC 값으로부터, **D**는 DMG와 DCLD 값으로부터 계산이 가능하다.

이러한 과정으로 구해진 **G**는 디코릴레이터의 적용을 고려하여 $\mathbf{G}_{Mod}$로 변환되어 사용된다.

3.3.3. SAOC TRANSCODER MODE

SAOC가 Transcoder Mode로 동작할 때의 구조는 그림 7과 같다. 그림에서 볼 수 있는 것처럼, SAOC Transcoder의 입력은 SAOC 인코더에서 생성된 다운믹스와 SAOC bitstream과 사용자가 정의한 Rendering Matrix이고, 출력은 전처리 된 다운믹스 신호와 MPEG Surround bitstream이다.

SAOC Decoder 모드와 비교하면, parameter processor가 MPEG Surround bitstream을 출력하고, Downmix processor가 최종 rendering된 오디오 신호가 아닌 MPEG Surround 입력이 되는 downmix 신호를 출력한다는 점이 크게 다르다. 이때 Downmixprocessor는 입력이 스테레오인 경우에만 동작하며, 모노인 경우는 bypass하게 된다.

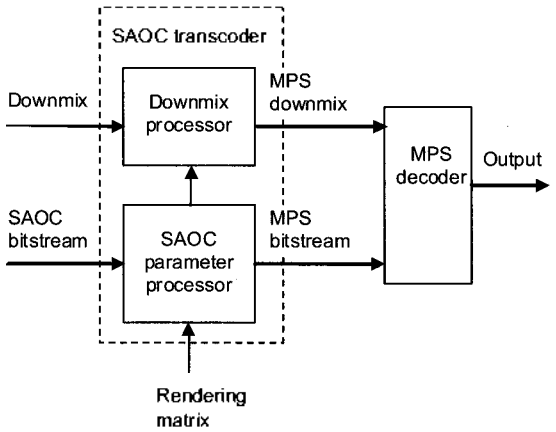

그림 7. SAOC Transcoder.
Figure 7. SAOC Transcoder.

사용자가 입력한 Rendering Matrix에 따른 sound scene은 MPEG Surround bitstream에 표현되어 담겨있고, 최종 rendering된 멀티 채널 출력은 MPEG Surround decoder가 담당하게 된다. Transcoding을 통해 생성되는 MPEG Surround 파라미터로는, 다운믹스 신호가 스테레오일 경우, TTT (Two-to-Three) box에 사용되는 prediction matrix와, 3개의 OTT (One-to-Two) box에 사용되는 CLD (Channel Level Difference) 값들이 있다. 예측 행렬인 $\mathbf{C}_{TTT}$는 아래 관계로부터 추정이 가능하고,

$$\mathbf{C}_{TTT}\mathbf{G} = \mathbf{C}_3 \tag{16}$$

각 CLD 값은 원하는 출력 신호의 공분산 행렬 **F**를 이용하여 아래와 같이 구할 수 있다.

$$\mathbf{F} = \mathbf{YY}^* = \mathbf{A}(\mathbf{SS}^*)\mathbf{A}^* = \mathbf{AEA}^* \tag{17}$$

$$\mathbf{F} = \begin{pmatrix} f_{11} & f_{12} & \cdots & \cdots & f_{15} \\ f_{21} & f_{22} & & & \vdots \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ f_{51} & \cdots & \cdots & \cdots & f_{55} \end{pmatrix} \tag{18}$$

$$CLD_{1,2} = 10\log\left(\frac{f_{11}}{f_{22}}\right) \tag{19}$$

$$ICC_{1,2} = \frac{f_{12}}{\sqrt{f_{11}f_{22}}} \tag{20}$$

만약 다운믹스 신호가 모노일 경우, TTT box를 사용하지 않으므로, 식 (17)~(20)의 과정을 이용하여 총 5개의 OTT box에 사용되는 CLD, ICC를 구하게 된다.

3.3.4. ENHANCED MODE

SAOC의 주요 application의 하나인 Interactive Remix의 경우, 핵심 rendering scene 중에 Karaoke (보컬 object를 완전히 제거) 혹은 Solo (보컬을 제외한 나머지 object를 모두 제거)가 있는데, SAOC의 parametric한 부가 정보만을 이용하여 이와 같은 극단적인 object 제어를 하는 경우 성능 열화가 크다. 즉, 완전히 제거를 하는 경우 남아있는 오디오 신호의 왜곡이 매우 심하거나, 왜곡을 최소화하기 위해 제거 (attenuation)을 일정 수준에서 제한해야만 하는 문제를 가지고 있었다. 이와 같은 문제를 해결하기 위해 MPEG Surround에서 사용되던 residual 신호의 bitstream 전송 개념을 SAOC에 도입하였다.

Karaoke를 위한 보컬 경우처럼 극단적인 제어를 요구하는 오브젝트에 대해서는 기본 부가 정보 이외에 오브젝트 신호 자체를 AAC 등의 waveform 코덱으로 부호화한 데이터 (residual)를 부가적으로 전송하고 이를 활용하여 SAOC rendering의 성능을 향상시키는 기술이 표준에 포함되었다. 이렇게 residual을 포함한 오브젝트를 EAO (Enhanced Audio Object)라고 하며, 그림 8은 EAO를 포함한 SAOC decoder/transcoder의 구조를 나타낸다. EAO를 포함한 object는 일반적인 SAOC와는 다른 decoding

과정을 거치게 되는데, 그림에서 Residual processor는 이를 수행하는 block이다. 또한 Residual processor에 의한 처리 과정 이외에 일반적인 오브젝트 파라미터를 이용한 처리과정은 SAOC downmix pre-processor에서 수행되며, EAO와 일반적인 오브젝트에 대한 처리 결과는 object combiner에 의해 합성되어 최종적인 출력신호를 얻을 수 있다.

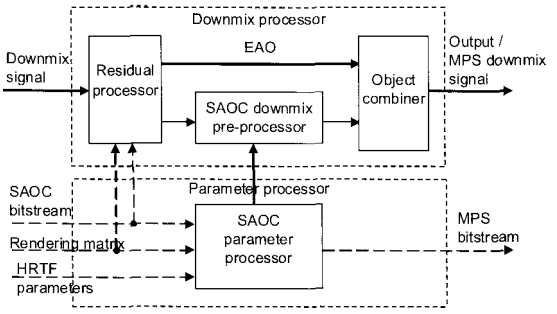

그림 8. Enhanced Audio Object (Residual)이 사용된 경우의 decoder 구조.
Figure 8. Decoder with Enhanced Audio Object (Residual).

3.4. MCU COMBINER

앞서 2.2.3에서 언급한 바와 같이, SAOC의 주요 application target 중 하나는 원격 회의이다. 기존의 원격 회의 시스템에서는 MCU (Multi-point Control Unit)를 이용하여 다지점에서 전송되는 음성 신호를 제어하여 각 수신단에 보내준다. SAOC를 이용하는 원격 회의의 경우, 다운 믹스 신호는 기존의 음성신호를 처리하는 방법과 동일하게 MCU를 통하여 송신/수신 할 수 있으나, SAOC bitstream에 대해서는 MCU에서 병합/전송을 하기 위한 방법이 새롭게 정의되어야 한다.

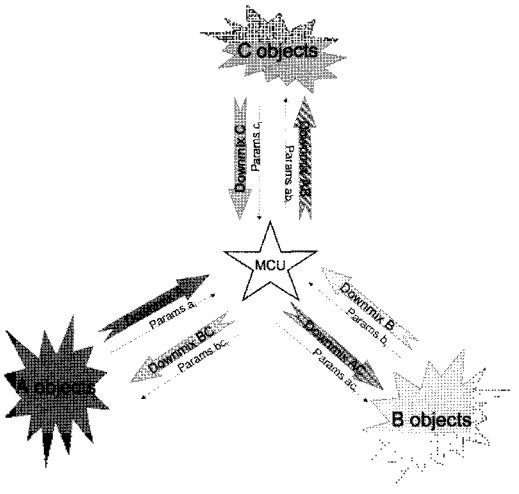

그림 9. Teleconference with MCU.
Figure 9. Teleconference with MCU.

그림 9에서 볼 수 있는 것 같이, A 지점에서는 A 지점에 위치한 화자들의 다운믹스 신호와 화자들에 대한 SAOC 파리미터를 MCU로 전송하고, MCU로부터 B 지점에 위치한 화자들의 다운믹스 신호와 C 지점에 위치한 화자들의 다운믹스 신호가 병합된 다운믹스 신호를 전송 받는다. 또한, SAOC rendering을 위해 필요한 B 지점의 화자들에 대한 파리미터와 C 지점의 화자들에 대한 파리미터를 전송받을 필요가 있는데, SAOC MCU Combiner가 이 역할을 담당한다. MCU Combiner의 구조는 그림 10과 같다.

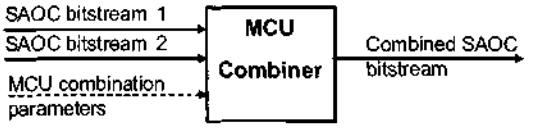

그림 10. MCU Combiner.
Figure 10. MCU Combiner.

효율적인 bitstream 병합을 위해, MCU Combiner에서는 병합되는 두 지점의 오브젝트 파리미터들을 다시 추정(디코딩 후 재 인코딩)하는 것이 아니라, MCU에 전송된 파라미터들에 대해 정규화 값만을 변형하여 병합하는 형태로 동작된다. 이때 변형되는 오브젝트 파리미터는 OLD와 NRG로, 다음과 같은 수식에 의해 변형된다.

$$OLD_j^{comb}(pb) = \frac{G^{comb} \cdot OLD_i^{bs1}(pb) NRG^{bs1}(pb)}{NRG^{comb}(pb)} + \frac{(1-G^{comb}) \cdot OLD_i^{bs2}(pb) NRG^{bs2}(pb)}{NRG^{comb}(pb)} \quad (21)$$

$$NRG^{comb}(pb) = \max(G^{comb} \cdot NRG^{bs1}(pb), (1-G^{comb}) \cdot NRG^{bs2}(pb)) \quad (22)$$

여기서, j는 병합된 bitstream의 오브젝트 인덱스이고, i는 병합될 bitstream의 오브젝트 인덱스이며, comb, bs1, bs2는 각각 병합된 비트스트림, 첫 번째 입력 비트스트림, 두 번째 비트스트림을 지칭하는 인덱스이다. SAOC 비트열에 포함된 파라미터들 가운데 NRG는 이와 같은 MCU Combining이 필요한 경우에만 사용되는 파라미터이며, 거꾸로 NRG가 없는 SAOC 비트열은 MCU Combiner를 통한 병합을 수행할 수 없다.

3.5. OTHER FUNCTIONALITIES

3.5.1. Preset

앞서 언급된 것과 같이 SAOC는 다른 Audio Codec들과

는 다르게 사용자로부터 직접 입력을 받는 Rendering matrix가 있어야만 복호화가 가능하다. 바꿔 말하면, 사용자 입력이 없거나 잘못된 Rendering matrix를 부여할 경우 복호화가 불가능하다. 이와 같은 특성은 SAOC로 하여금 오디오 신호 자체에 해당하는 파라미터의 전송 및 이를 복호화하는 방법에 대한 기술 이외의 다른 부분들을 표준화하는데 이르게 하였다. 대표적인 것이 Rendering matrix를 전송하는 Preset이다. Preset은 Rendering matrix 자체를 효율적으로 전송, 저장하기 위한 syntax로써, Interactive Re-mix에서는 컨텐츠 제공자가 직접 몇 개의 추천하는 rendering scene (예를 들어, Karaoke, Club mix, Acoustic mix, 등)을 SAOC 비트열과 함께 전송하는데 사용할 수 있고, Teleconference에서는 MCU가 원격의 회의 참여자들을 가상 공간에 위치시키기 위한 정보의 전송에 활용 가능하다. 또한, 사용자가 본인만의 최적의 Rendering scene으로 design한 Preset 정보를 제3자에게 전송하는 형태로도 활용이 가능하다. 단, 이 경우 Preset은 SAOC가 아닌 독립된 파일 혹은 비트열의 형태로 저장/전송된다. Preset은 하나의 SAOC 비트열에 대해 고정된 값을 갖는 Static preset과 시간에 따라 가변할 수 있는 Dynamic preset으로 구별된다.

Preset과 더불어 사용자로부터 올바른 Rendering matrix를 입력받기 위해서는 전송된 object의 개수와 함께 해당 object가 무엇인지를 알려주는 정보가 필요한데 이를 위한 metadata도 SAOC 비트열 syntax에 정의되어 있다. 또한, FAO가 아닌 오브젝트의 경우 극단적인 rendering 요구는 음질 성능을 크게 저하시키는 문제를 발생시키게 되는데, 이에 대한 보호를 위해 Rendering matrix의 범위를 제한하기 위한 가이드 정보를 전송하는 방법에 대해서도 현재 표준화 논의가 되고 있다.

3.5.2. Low power mode

SAOC 복호화를 위해서는 입력신호를 Hybrid QMF sub-band domain으로 변환하는 과정이 필요한데, 이는 복소수 filterbank 형태이기 때문에 구현시 연산량이 매우 높은 문제를 가진다. MPEG Surround에서는 일부 밴드에 대해 실수 연산 filterbank를 이용하고, decorrelator의 연산량도 낮추어 수행하는 형태의 decoding mode를 별도로 두어 decoder 구현을 선택적으로 하는 전략을 취했다. 이와 같은 Low power mode는 약간의 음질 저하를 감수하면, 연산량을 50%까지 낮출 수 있다. SAOC에서도 연산량 감소 및 MPEG Surround의 Low power mode와의 호환성을 위해 비슷한 방법으로 Low power mode를 정의하였다.

3.5.3. Low delay mode

Teleconference에 SAOC를 이용하기 위해서는 양방향 통신에 필요한 만큼의 encoding-decoding delay 요구사항 (통상 50 msec 이내)을 만족해야 한다. SAOC에서 사용하는 Hybrid QMF filterbank는 이와 같은 요구조건을 만족시키지 못하기 때문에, Low delay 구현을 위해 별도의 mode를 정의하고 적합한 filterbank 및 그 밖의 low delay를 위한 제약사항을 정의하고 있다. 그러나 MPEG Surround는 Low delay decoding mode가 정의되어 있지 않기 때문에, SAOC가 transcoding 모드로 동작하는 경우에 대한 Low delay 구현에 어려움이 있는데, 현재 MPEG Surround에 대한 low delay 구현 방안에 대해서도 표준 회의에서 논의가 진행되고 있다.

IV. 결 론

MPEG SAOC는 PS, MPEG Surround로 이어져온 MPEG 오디오의 parametric 부호화 기술을 계승 발전시킨 표준 기술로써, 특히 지금까지의 채널 기반 오디오 부호화의 사고에서 MPEG-4의 근본 사상이던 객체 기반 부호화로의 의미 있는 전환을 시도한 기술이라는데 큰 의미를 부여할 수 있다. 이와 같은 객체 기반 오디오 부호화 기술은 Interactive Re-mix, 방송에서의 Flexible object rendering, Immersive teleconference 등 여러 응용 분야의 차세대 오디오 기술로써 활용될 수 있다. 현재 MPEG SAOC는 FCD 단계를 지나 표준화의 최종 단계인 FDIS를 예정해 두고 있다. 객체 기반, 사용자 입력, Transcoding 등 오디오 표준화에서 처음 시도되는 개념들로 인해, 성능 평가 방법 등 많은 부분에서 기존의 표준화 paradigm과 다른 이슈들을 풀어오면서 표준화가 이뤄졌고, 그 사례들은 이후 다른 표준화 및 연구 과제에서 활용될 수 있을 것이다.

참 고 문 헌

1. ISO/IEC Int. Std. 11172-3:1993, *Information technology - Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbit/s - Part 3: Audio*,, 1993.
2. ISO/IEC Int. Std. 13818-7:1997, *Information technology - Generic coding of moving pictures and associated audio information - Part 7:Advanced Audio Coding (AAC)*, 1997.
3. ISO/IEC Int. Std. 14496-3:1999, *Information technology - Coding of audio-visual objects - Part 3: Audio*, 1999.
4. ISO/IEC Int. Std. 14496-3, Amd.1:2003, *Information technology*

- *Coding of audio-visual objects – Part 3: Audio*, 2003.
5. ISO/IEC Int. Std. 14496-3, Amd.2:2004, *Information technology – Coding of audio-visual objects – Part 3: Audio*, 2004.
6. http://www.ebu.ch/CMSimages/en/tec_doc_t3296_tcm6-10497.pdf
7. ISO/IEC Int. Std. 23003-1:2007, *MPEG audio technologies - Part 1: MPEG Surround*, 2007.
8. F. Baumgarte and C. Faller, "Binaural Cue Coding – Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.
9. C. Faller and F. Baumgarte, "Binaural Cue Coding – Part II: Schemes and applications," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.
10. ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. M13632, "From Channel-Oriented to Object-Oriented Spatial Audio Coding," Klagenfurt, July 2006.
11. ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. N8853, "Final Call for Proposals on Spatial Audio Object Coding," Morocco, January 2007.
12. ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. N9250, "Report on Spatial Audio Object Coding RM0 Selection," Lausanne, July 2007.
13. ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. N13843, "ISO/IEC FCD 23003-2:200x, Spatial Audio Object Coding," London, July 2009.
14. T. Lee, J-H. Yoo, and D. Jang, "A Personalized preset-based audio system for interactive service," in *AES 121st convention*, preprint 6904, San Francisco, USA, Oct., 2006.
15. ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. M14159, "Comments on Draft Call for Proposals on Spatial Audio Object Coding," Marrakech, Jan 2007.

저자 약력

•정 양 원 (Yang-Won Jung)

1998년 : 연세대학교 전자공학과 (학사)
2000년 : 연세대학교 전기컴퓨터공학과 (석사)
2005년 : 연세대학교 전기전자공학과 (박사)
2005년~2009년 : LG전자 Digital TV연구소 책임연구원
2009년~현재 : 인텔렉추얼 벤처스 코리아 부장
※주관심 분야 : 오디오 신호처리, 오디오/음성 코덱 표준화

•오 현 오 (Hyen-O Oh)

1996년 : 연세대학교 전자공학과 (학사)
1998년 : 연세대학교 전자공학과 (석사)
2002년 : 연세대학교 전기전자공학과 (박사)
2002년~현재 : LG전자 Digital TV연구소 책임연구원
※주관심 분야 : 오디오 신호처리, 오디오/음성 코덱 표준화