한국음향학회지 제32권 제3호 pp. 227~236 (2013)
The Journal of the Acoustical Society of Korea Vol.32, No.3 (2013)
http://dx.doi.org/10.7776/ASK.2013.32.3.227

pISSN : 1225-4428
eISSN : 2287-3775

교차 예측 기반의 보컬 추정 방법을 이용한 SAOC Karaoke 모드에서의 음질 향상 기법에 대한 연구

Quality Improvement of Karaoke Mode in SAOC using Cross Prediction based Vocal Estimation Method

이동금†, 박영철*, 윤대희

(Tung Chin Lee, Young-Cheol Park*, and Dae Hee Youn)

연세대학교 전기전자공학과, *연세대학교 컴퓨터정보통신공학부

(접수일자: 2012년 12월 26일; 수정일자: 2013년 2월 13일; 채택일자: 2013년 2월 18일)

초 록: 본 논문에서는 SAOC의 Karaoke 모드의 출력 신호 내에 존재하는 잔여 보컬 성분을 추정하여 억제시킴으로써 음질을 향상시킬 수 있는 알고리듬을 제안하였다. 잔여 보컬 성분은 Karaoke 모드 환경으로 합성된 신호와 Solo 모드로 새로 합성된 신호를 서로 교차 예측하여 추정될 수 있다. 그러나, 두 신호는 모두 같은 다운 믹스 신호로부터 합성되는 신호이므로, 두 신호간의 높은 상관성으로 인하여 가라오케 신호내의 잔여 보컬 성분뿐만 아니라 음악 성분도 함께 제거된다. 이러한 열화를 해결하기 위해, 본 논문에서는 교차 예측 과정에서 심리 음향적 특성을 고려한 예측 방해 신호를 적용하였으며, 이 신호의 크기는 심리음향모델의 마스킹 특성에 따라 음악적 음질의 열화가 최소화되도록 적응적으로 설정되었다. 실험은 보컬 객체가 포함된 음악 신호에 대해서 객관적 및 주관적 음질평가를 수행하였으며, 전체적으로 성능 향상이 있음을 확인하였다.

핵심용어: Spatial Audio Object Coding (SAOC), Karaoke, 교차 예측, 보컬 추정

ABSTRACT: In this paper, we present a vocal suppression algorithm that can enhance the quality of music signal coded using Spatial Audio Object Coding (SAOC) in Karaoke mode. The residual vocal component in the coded music signal is estimated by using a cross prediction method in which the music signal coded in Karaoke mode is used as the primary input and the vocal signal coded in Solo mode is used as a reference. However, the signals are extracted from the same downmix signal and highly correlated, so that the music signal can be severely damaged by the cross prediction. To prevent this, a psycho-acoustic disturbance rule is proposed, in which the level of disturbance to the reference input of the cross prediction filter is adapted according to the auditory masking property. Objective and subjective test were performed and the results confirm that the proposed algorithm offers improved quality.

Keywords: Spatial audio object coding (SAOC), Karaoke, Cross prediction, Vocal estimation
PACS numbers: 43.60. Ek

I. 서 론

최근 들어, 다양한 어플리케이션 (Home Theater, Movie, Game, etc.)에서 오디오 재생 시스템을 통해 보다 높은 공간감 및 현장감을 구현하기 위해 5.1채널 혹은 그 이상의 채널 스피커들이 사용되고 있다. 하지만 멀티채널 스피커 환경을 재생하기 위해서는 스피커 증가에 비례해서 데이터 전송률도 비례하여 함께 증가하기 때문에, 대역폭이 제한되는 방송 혹은 양방향 통신 환경에서는 구현하기 어렵다. 이러한 제한 요소를 극복하기 위해 사람의 청각적인 특성을 이용하는 "공간 오디오 부호화(Spatial Audio Coding)" 기술이 연구되었고, MPEG에서는 이 기술을 기초로 해서 "공간 오디오 객체 부호화(SAOC,

†**Corresponding author:** Tung Chin Lee (riglord@dsp.yonsei.ac.kr)
B601, 2nd Engineering Building, Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul 120-749, Republic of Korea.
(Tel: 82-2-2123-4534, Fax: 82-2-364-4870)

Spatial Audio Object Coding)" 기술을 표준화하였다.[1] SAOC는 채널 기반이 아닌 오디오 객체 기반으로 처리되기 때문에, 수신단에서 사용자 목적에 따라 이득(gain) 이나 공간적 위치(spatial position) 등을 임의로 설정할 수 있는 렌더링(rendering) 정보가 사용된다. 이 정보를 통해 SAOC는 어떤 스피커 환경에도 적응적으로 신호를 재생할 수 있다.

SAOC는 입력 객체들을 모노(mono) 혹은 스테레오(stereo) 로 다운믹스(downmix) 하고, 각 객체의 특성을 파라미터(parameter)로 변환하여 전송하기 때문에, 수신단에서 입력된 객체의 파형(waveform)을 완벽하게 복원하기 어렵다.[1-2] 특히, Karaoke 모드와 같이 보컬(vocal) 객체를 완벽하게 제거 시켜야하는 환경에서는 다운 믹스된 신호로부터 여러 객체들이 완벽하게 분리되지 않는 한계 때문에, 합성된 신호에 보컬 특성이 함께 섞여서 출력된다. 이러한 약점을 극복하기 위해 SAOC 에서는 부호화단에서 파라미터 부호화를 통해 발생되는 오차 신호(error signal)를 추가로 전송하여, 합성된 신호로부터 발생할 수 있는 음질적 열화를 보상해주는 잔여 부호화(residual coding)방법을 사용 한다.[3] 잔여 부호화를 사용하지 않는 환경에서도 부호화단에서 보컬 객체의 기초 주파수(fundamental frequency) 정보를 추정하여 전송함으로써 복호화 과정에서 합성되는 하모닉(harmonics) 성분들을 제거하는 방법을 사용하여 복호화의 성능을 향상시킬 수 있다.[4] 하지만, 이 방법의 경우 DFT(Discrete Fourier Transform) 대역에서 연산이 이루어지기 때문에, 해당 알고리듬을 SAOC 표준 비트스트림(bitstream)에 직접적으로 적용할 수 없는 단점이 있다.

본 논문에서는 SAOC 표준 비트스트림을 이용한 환경에서 Karaoke 모드의 전체적인 음질을 향상시킬 수 있는 알고리듬을 제안하였다. 제안된 알고리듬의 주 목적은 부호화단에서 추가정보 없이 Karaoke 모드로 복호화된 뒤 남아 있는 보컬 성분들을 제거하여 전체적인 음질을 향상시키는 방법이다. 이는 Karaoke 모드로 복호화된 신호와 SAOC의 솔로(Solo) 모드로 추가로 복호화된 신호를 교차 예측(cross prediction)을 수행함으로써 Karaoke 신호에 남아있는 보컬 성분을 추정하여 제거 할 수 있다. 하지만, 교차 예측의 입력으로 사용되는 두 신호가 똑같은 다운믹스 신호로부터 복호화되기 때문에 두 신호간의 상관성이 매우 높다. 이는 보컬 성분뿐만 아니라, 음악 성분(보컬이 아닌 성분)도 함께 예측을 하여 음악 신호의 음질 열화도 함께 초래한다. 이러한 문제를 보완하기 위해, 본 논문에서는 교차 예측 과정에서 예측을 억제하기 위해 심리 음향적 특성을 고려한 방해 신호(disturbance signal)을 적용하였으며, 이 방해 신호의 크기는 사람의 청각적인 특성을 적용하여 Karaoke 음질의 열화를 최소화하며 보컬 성분을 추정하도록 설정되었다. 그리고 객관적, 주관적인 실험 결과를 통해 제안된 알고리듬이 효과적이었음을 확인하였다.

본 논문은 2장에서는 SAOC의 기본 구조를 간단히 소개하고 Karaoke 모드에서의 문제점을 지적하였으며, 3장에서는 본 논문에서 제안하는 교차 예측을 기반으로 한 보컬 추정 알고리듬을 설명한다. 4장에서는 제안된 알고리듬의 성능을 평가하기 위해 객관적 및 주관적 실험 결과를 보여주고 마지막 5장에서 결론을 맺는다.

II. SAOC의 기본적인 구조와 Karaoke 모드에서의 문제

기본적인 SAOC의 부호화기와 복호화기의 구조는 Fig. 1 과 같다.

Fig. 1(a)의 부호화기에서 입력되는 오디오 객체들

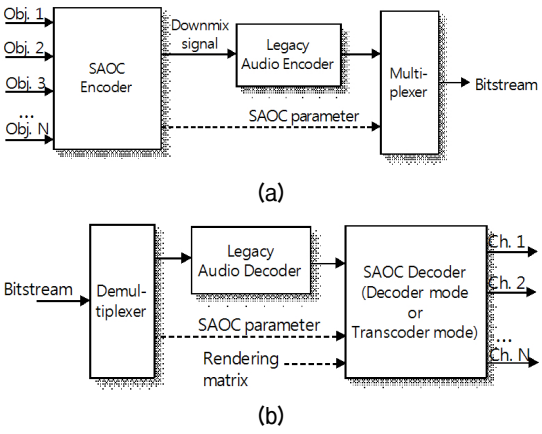

Fig. 1. Spatial Audio Object Coding (a) Encoder (b) Decoder.

은 hybrid-QMF 필터뱅크(filterbank)를 통해 시간-주파수 신호로 변환된다. 다음, OLD(Objective Level Difference), IOC(Inter Object Correlation)와 DMG(DownMix Gain), 등과 같은 공간 객체 파라미터들을 각 주파수 밴드 단위로 추출한다. 입력 오디오 객체 신호들은 모노 혹은 스테레오 신호로 다운믹스 되고, 현존하는 오디오 부호화기(AAC, MP3 등)를 이용하여 부호화되어, 추출된 파라미터와 함께 비트스트림으로 전송된다.[5]

Fig. 1(b)의 SAOC 복호화기는 비트스트림으로부터 전송된 다운믹스 신호와 공간 객체 파라미터뿐만 아니라, 사용자 인터페이스 기반의 렌더링 매트릭스[5]도 함께 이용하여 출력 채널 환경에 맞는 공간 파라미터를 계산한다. 다음, 이 공간 파라미터를 다운 믹스 신호에 적용하여 원하는 채널 환경에 맞게 합성한다. 이 때, SAOC 복호화기의 연산량이 효율적으로 사용되도록 출력하고자하는 채널 수에 따라 디코더 모드(decoder mode)와 트랜스코더 모드(transcoder mode)로 구분된다. 예로, 모노, 스테레오, 바이노럴 스테레오(binaural stereo) 신호를 출력할 때는 디코더 모드가 동작하며, 그 이상의 채널(5.1 채널) 신호들을 출력할 때에는 트랜스코더 모드로 동작한다. 트랜스코더 모드에서는 먼저 입력 다운믹스 신호와 객체 파라미터들을 MPEG 서라운드(MPEG Surround(MPS))[6] 복호화기에 적합한 비트스트림으로 변환되고, 최종 멀티채널 신호는 MPS 복호화기를 통해 출력된다.

SAOC의 렌더링 매트릭스는 사용자에 의해 완전히 제어된다. 만약 N개의 객체가 입력으로 사용된다면, 임의의 주파수 밴드 b에서의 출력 신호의 파워는 다음과 같이 계산된다.

$$\begin{aligned} G_C^2 &= \sum_{i=0}^{N-1}\left(\sum_{j=0}^{N-1} m_{i,C} m_{j,C} e_{i,j}[b]\right) \\ &= \sum_{i=0}^{N-1}((m_{i,C} m_{0,C}) e_{i,0}[b]) + \ldots + (m_{i,C} m_{N-1,C}) e_{i,N-1}[b]). \end{aligned} \quad (1)$$

식에서 $e_{i,j}[b]$는 주파수 밴드 b에서 입력 오디오 객체 i와 j 사이의 공분산(covariance)을 나타내며, $e_{i,j}[b] = \sqrt{OLD_i[b] OLD_j[b]}\, IOC_{i,j}[b]$로 정의될 수 있다. 식에서 IOC와 OLD 파라미터들은 각각 $OLD_i[b] = nrg_{i,i}[b]/\max(nrg_{i,i}[b])$, $i = 1,...,N$와 $IOC_i[b] = nrg_{i,j}[b]/\sqrt{nrg_{i,i}[b] nrg_{j,j}[b]}$로 나타낼 수 있으며, $nrg_{i,j}[b]$는 밴드 b에서 객체 i와 객체 j와의 곱을 의미한다. 렌더링 매트릭스의 원소(element) $m_{k,l}$는 k번째 객체를 l번째 채널에 할당하는 게인 값을 나타낸다. 그러므로 입력 N개의 오디오 객체가 모두 서로 비상관(decorrelated) 되어 있고, N번째 객체가 보컬 객체라고 가정한다면, Karaoke 모드로 복호화 되는 식(1)은 아래 식으로 간략화될 수 있다.

$$G_C^2(b) = \sum_{i=0}^{N-1}(m_{i,C} m_{i,C}) e_{i,C}(b). \quad (2)$$

다운 믹스 신호는 입력 오디오 객체들의 합으로 구성되기 때문에, 모든 오디오 객체들의 공간 단서(spatial cue)들은 다운 믹스 신호 내에 섞여있다. 이 때문에 음질의 왜곡 없이 특정 단서를 완전히 제거하기가 불가능하므로, Karaoke 환경과 같이 특정 객체만을 완전히 제거하는 분야에 사용하기에는 적합하지 않다. 그러므로, 본 논문에서는 후처리 연산을 통해 음질의 왜곡을 최소화하며 특정 객체의 영향을 줄여줄 수 있는 방법(본 논문에서는 Karaoke 모드에서 보컬 객체의 영향을 최소화)을 제안하였다.

III. SAOC에서의 보컬 추정 알고리듬

SAOC의 복호화가 잘 이루어질 수 있다면, 이상적으로는 Karaoke 모드에서 음악(보컬 객체만 제거된) 신호만을 복원할 수 있어야한다. 하지만, 실제로는 입력 객체들 간의 분리가 완벽하지 않기 때문에, Karaoke 모드에서 복원된 음악 신호에는 언제나 보컬 객체 성분이 섞여있다. 이러한 원치 않는 보컬 성분들은 예측 필터를 이용하여 제거할 수 있다. 기본적인 원리는 SAOC의 솔로 모드(보컬 객체만 합성)로부터 새로 합성된 신호를 Karaoke 모드 신호와 함께 예측 필터기의 입력으로 사용하여, Karaoke 신호 내에 존재하는 원치 않는 보컬 성분들을 추정하여, 그 성분을 제거하는 것이다. 하지만, 솔로 모드로부터 합성된 신호 역시 마찬가지로 완벽히 분리되지 않기 때문에 음악 객체 성분들이 보컬 신호에 섞여 있게 된다. 그러므로 Karaoke모드와 솔로 모드를 통해 만들어진 두 신호는 전대역에 대해서 상당히 높

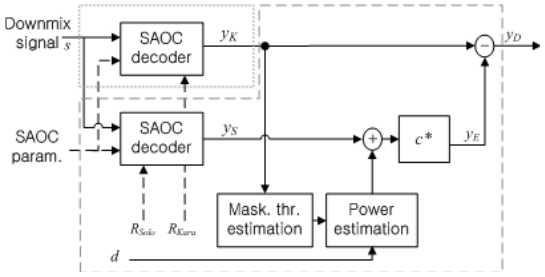

Fig. 2. Block diagram of the proposed vocla estimation.

은 상관성을 가지게 된다. 이러한 경우에 대해서 예측 필터를 이용한다면 음악 신호내의 보컬 성분뿐만 아니라, 음악 성분 신호도 함께 추정하게 되는 문제가 발생하기 때문에, 예측 필터를 그대로 사용할 수 없다. 본 논문에서는, 심리음향 모델의 마스킹 특성을 이용하여 음악 성분들에게는 청각적으로 변화를 주지 않는 예측 필터를 설계하여 음악 신호 내에 존재하는 보컬 성분을 추정하여 제거하는 알고리듬을 제안하였다.

아래부터는 설명의 편의성을 위해 입력되는 여러 객체들을 음악과 보컬 두 객체만으로 구분하였다. Fig. 2는 Karaoke 신호 내에 존재하는 잔여 보컬을 제거하기 위한 전체적인 블록도를 도시하였다. Fig. 2에서 점선 영역은 일반적인 SAOC 복호화 과정을, 파선 영역은 새로 제안된 알고리듬을 나타내었다. 그림에서 나타나 있듯이 잔여 보컬은 Karaoke 모드로 복호화된 신호 y_K와 솔로 모드로 복호화된 신호 y_S와의 교차 예측을 수행하여 추정하는데, 이 때 사용된 최적의 필터 계수 c는 최소 평균 제곱 에러(Minimum Mean Square Error) 척도를 이용하여 계산하였다. 그리고 음악 신호의 열화를 방지하게 위해 마스킹 특성이 반영된 예측 방해 신호 d를 예측 과정에 추가하였다. 이에 대해서는 뒤에서 자세히 언급한다.

솔로 신호 y_S는 복호화 과정에서 새로 입력되는 렌더링 매트릭스 R_{Solo}를 통해 복호화 되는데, 이 매트릭스 내의 원소는 Karaoke 모드로 사용되는 렌더링 매트릭스 R_{Kara}와 정 반대로 구성된다. 예로, 만약 입력으로 두 객체(첫 번째 객체는 음악 객체, 두 번째 객체는 보컬 객체)가 사용되었다고 가정하면, 두 렌더링 매트릭스 R_{Kara}와 R_{Solo}는 각각 다음과 같이 구성된다, $R_{Kara} = [Obj_{Nonvocal}\ Obj_{Vocal}] = [1\ 0]$ and $R_{Solo} = [0\ 1]$.

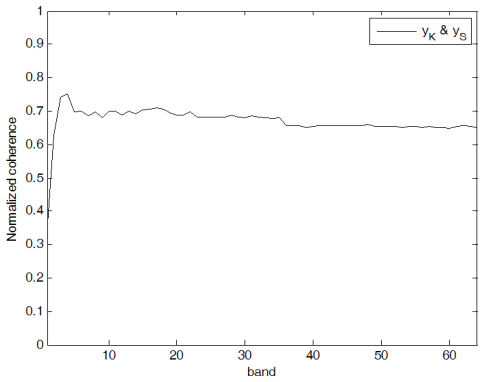

Fig. 3. Correlation between y_K and y_S.

입력으로 사용된 객체들은 복호화 과정에서 완벽히 분리되지 않기 때문에, Karaoke 신호 y_K와 솔로신호 y_S 신호 내에는 각각 잔여 보컬 성분과 잔여 음악 성분이 입력 객체 파워의 크기에 따라 주파수 대역별로 서로 다른 레벨로 섞이게 되어, 두 신호는 전대역에 대해서 서로 높은 상관성을 가지게 된다. Fig. 3은 임의의 보컬과 음악 객체들로 구성된 신호에 대해서 y_K와 y_S 사이의 상관도를 나타낸 것이다.

Fig. 3에서 가로 축은 주파수 밴드, 세로축은 아래와 같이 각 주파수 밴드단위로 정의되는 정규화된 상관 관계(normalized coherence)를 나타내었다.

$$\rho = \frac{E\left[Y_K[k]\, Y_S^*[k] \right]}{\sqrt{\sigma_K^2 \sigma_S^2}}. \tag{3}$$

식에서 $\sigma_K^2[k]$와 $\sigma_S^2[k]$는 각각 y_K와 y_S의 파워를, $Y_K[k]$와 $Y_S[k]$는 각각 y_K와 y_S의 주파수 계수값을 의미한다. 그림에서 확인할 수 있듯이, 두 신호간의 상관 관계가 높아, 교차 예측기를 그대로 적용하면, y_K 신호 내에 있는 보컬 성분뿐만 아니라, 음악 성분도 함께 추정되어 제거될 수 있다. 만약 예측 방해 신호 d를 적용하지 않고, 교차 예측을 수행하면 추정된 신호 y_C는 주파수 축에서 아래와 같이 나타낼 수 있다.

$$Y_C[k] = Y_K[k] - C^* Y_S[k]. \tag{4}$$

식에서 C는 예측 계수를, *은 복소 켤레를 의미한다. 최적의 예측 계수는 최소 평균 제곱 에러 척도를

이용하여 다음 식으로 계산된다.

$$C^o = \frac{E\left[Y_K[k]\, Y_S^*[k] \right]}{E\left[|Y_S[k]|^2 \right]}, \tag{5}$$

식(5)에 식(3)을 적용하면, 아래 식으로 다시 표현될 수 있다.

$$C^o = \rho \sqrt{\frac{\sigma_K^2}{\sigma_S^2}}. \tag{6}$$

이 결과를 이용하면 추정된 신호의 파워는 다음 식으로 계산된다.

$$\begin{aligned} E\left[|Y_C[k]|^2 \right] &= \sigma_{y_C}^2 = E\left[|Y_K[k] - C^o Y_S[k]|^2 \right] \\ &= \sigma_K^2 + \left(|\rho| \sqrt{\frac{\sigma_K^2}{\sigma_S^2}} \right)^2 \sigma_S^2 - 2|\rho|^2 \sqrt{\frac{\sigma_K^2}{\sigma_S^2}} \sqrt{\sigma_K^2 \sigma_S^2} \\ &= (1 - |\rho|^2)\sigma_K^2. \end{aligned} \tag{7}$$

식(7)로부터 추정된 신호의 파워는 $(1-|\rho|^2)$의 인수 값으로 억제되는 것을 알 수 있다. 하지만, 솔로 신호 y_S 내에 존재하는 잔여 음악 신호로 인해, 직접적인 교차 예측 수행을 하게 되면 Karaoke 신호 내의 음악 성분도 함께 억제 시키는 문제가 발생한다. 예로, Fig. 3의 경우처럼 $\rho \approx 0.7$일 경우, 오직 30%에 해당되는 음악 신호의 파워만이 추정된 신호에 남아있게 된다. 이러한 문제를 해결하기 위해, 본 논문에서는 Fig. 2에 나타낸 것처럼 y_S 신호에 예측 방해 신호 d를 추가하였다. 신호 d는 평균이 0, 분산이 σ_D^2인 가우시안(Gaussian) 특성을 갖는 랜덤 신호로 가정하였다. 이러한 환경에서 교차 예측을 통해 추정된 출력 신호는 다음과 같다.

$$Y_D[k] = Y_K[k] - C_D^*\left(Y_S[k] + D[k] \right). \tag{8}$$

식에서 $D[k]$는 신호 d의 주파수 계수값을 의미한다. 이 때, Karaoke 신호와 솔로 신호 사이의 정규화된 상관 관계와 최적의 예측 계수를 계산하면 다음과 같이 나타낼 수 있다.

$$\begin{aligned} &\rho_D = \frac{E\left[Y_K Y_S^* \right]}{\sqrt{\sigma_K^2 \left(\sigma_S^2 + \sigma_D^2 \right)}} = \rho \sqrt{A}, \\ &where\ A = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_D^2}. \end{aligned} \tag{9}$$

$$C_D^o = \rho_D \sqrt{\frac{\sigma_S^2}{\sigma_S^2 + \sigma_D^2}}. \tag{10}$$

그러므로, 방해 신호가 추가된 환경에서의 추정된 신호의 파워는 아래와 같이 계산될 수 있다.

$$\begin{aligned} E\left[|Y_D[k]|^2 \right] &= \sigma_K^2 + \left(|\rho_D| \sqrt{A} \right)^2 \left(\sigma_S^2 + \sigma_D^2 \right) \\ &\quad - 2|\rho_D|^2 \sqrt{A} \sqrt{\sigma_K^2 \left(\sigma_S^2 + \sigma_D^2 \right)} \\ &= \left(1 - |\rho_D|^2 \right) \sigma_K^2 \\ &= \left(1 - |\rho|^2 A \right) \sigma_K^2. \end{aligned} \tag{11}$$

식(11)에서 확인할 수 있듯이, 추정된 신호는 σ_D^2에 의해 영향을 받는다. 그러므로 σ_D^2의 크기는 추정된 신호 내의 음악 성분에 영향이 가지 않도록 조절되어야 한다.

비록 SAOC의 복호화기가 다운 믹스 신호 내에 섞여 있는 객체를 완벽하게 분리하지 못한 채 출력 신호로 합성하지만, 부호화단에 입력되는 두 객체는 서로 독립적인 특성을 가질 수 있다. 이러한 경우, 복호화단에서 Karaoke 신호와 솔로 신호는 식(1)을 통해 다음과 같이 나타낼 수 있다.

$$\begin{aligned} Y_K[k] &= G_K^2[b] M[k] \\ Y_S[k] &= G_S^2[b] M[k]. \end{aligned} \tag{12}$$

식에서 $G_K[b]$과 $G_S[b]$는 각각 주파수 밴드 b에서의 음악 객체와 보컬 객체의 게인을, $M[k]$는 다운 믹스 신호를 의미한다. SAOC 표준 부호화기는 QMF 도메인 기반으로 이루어지므로, 밴드 단위로 신호의 파워를 계산하면 아래와 같이 나타낼 수 있다.

$$\begin{aligned} \sigma_K^2[b] &= G_K^2[b] \sigma_M^2[b] \\ \sigma_S^2[b] &= G_S^2[b] \sigma_M^2[b] \qquad b = 1, \cdots, B, \end{aligned} \tag{13}$$

에너지의 보존 법칙에 따라 두 신호 파워의 합은 다운믹스 파워의 합과 같아야 하므로 ,

$$\sigma_M^2[b] = \sigma_K^2[b] + \sigma_S^2[b], \tag{14}$$

식(14)와 $G_K^2[b] + G_S^2[b] = 1$의 조건을 식(13)에 대입하여 전개하면 아래 식으로 나타낼 수 있다.

$$\begin{aligned}\sigma_K^2[b] &= G_K^2[b]\sigma_K^2[b] + G_K^2[b]\sigma_S^2[b] \\ &= G_K^2[b]\sigma_K^2[b] + G_S^2[b]\sigma_K^2[b] \\ &= \sigma_{K\_M}^2[b] + \sigma_{K\_V}^2[b].\end{aligned} \tag{15}$$

윗 식은 Karaoke 신호 내에 있는 각 객체의 파워를 나타낸 것으로 생각 될 수 있다. 비록, SAOC의 Karaoke 신호 내에 섞여 있는 보컬의 성분 및 특성을 정확하게 추출할 수는 없지만, 출력 신호 내에 있는 파워 크기에 대해서는 에너지 보존 법칙에 의해 식(15)에서처럼 분리시켜 생각할 수 있다. 식(15)에서는 Karaoke 신호 내에 있는 음악 성분의 파워를 $\sigma_{K\_M}^2[b]$, 보컬 성분의 파워를 $\sigma_{K_V}^2[b]$로 나타내었다. 이 결과를 식(11)에 적용하면 추정된 신호의 파워는 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\sigma_{y_D}^2[b] &= (1-|\rho[b]|^2 A[b])\sigma_K^2[b] \\ &= (1-|\rho[b]|^2 A[b])(\sigma_{K\_M}^2[b] + \sigma_{K\_V}^2[b]) \\ &= \sigma_{y_{D^-}M}^2[b] + \sigma_{y_{D^-}V}^2[b]\end{aligned} \tag{16}$$

$\sigma_{y_{D^-}M}^2[b]$와 $\sigma_{y_{D^-}V}^2[b]$는 각각 음악 성분과 보컬 성분의 파워를 의미한다. 식(16)에서 확인할 수 있듯이, $\rho[b]$에 존재하는 방해 신호의 파워 $\sigma_D^2[b] = 0$, $|\rho[b]| = 1$이라면 $\sigma_{y_D}^2[b]$는 0이 되고, $0 < |\rho[b]| < 1$인 경우에는 음악과 보컬 성분이 동시에 제거된다. 더욱이 식(15)의 $G_K^2[b]$가 $G_S^2[b]$보다 크고, $|\rho[b]| \approx 1$이면 추정된 신호 내의 음악성분은 과도하게 제거된다. 그러므로 보컬을 추정하여 제거하는 과정에서 $\sigma_D^2[b]$의 크기는 음악의 음질을 열화시키지 않도록 적절히 조절되어야한다. 본 논문에서는 $\sigma_D^2[b]$를 계산하는 과정에서 사람의 청각적 마스킹 특성을 적용하였다. 즉, 음악 신호가 보컬 신호의 마스킹 임계치 아래에 존재하도록 추정 신호를 계산함으로써 음악의 청각적인 영향을 최소화시키는 것이다.

일반적으로 마스킹 임계치는 사람의 청각 구조의 주파수 선택도와 마스킹 특성을 통해 모델링된다[7]. 그러나 오디오 코딩 분야에서와 같이 정교한 계산이 필요한 것과는 달리 각 주파수 대역 별로 대략적인 마스킹 레벨만 있어도 충분하다고 판단되므로 본 논문에서는 식(17)의 상대적 임계 오프셋(relative threshold offset)[7]만 고려하여 전체적인 마스킹 임계치로 간주하였다.

$$O[b] = \alpha(14.5 + z) + (1-\alpha)5.5. \tag{17}$$

식에서 α는 토널리티(tonality) 인덱스를, z는 바크(bark) 스케일 주파수 인덱스를 의미한다. 정확한 보컬과 음악 성분의 마스킹 임계치는 보컬과 음악 신호로부터 계산되어야 하지만, 복호화된 신호내의 각 성분들은 정확히 분리되기 힘들다. 그러므로 주파수 밴드 단위로 각 성분들의 파워를 나눠서 표현한 식(15)를 이용하여 마스킹 임계치를 계산하였다. 예로, 보컬 성분의 마스킹 임계치는 식(18)에 표현된 것 같이 식(17)을 식(15)의 $\sigma_{K_V}^2[b]$에 적용하여 얻을 수 있다.

$$T_{K\_V}[b] = \frac{\sigma_{K\_V}^2[b]}{O[b]}. \tag{18}$$

이처럼, 본 논문에서 계산된 각 성분별 마스킹 임계치는 각 대역별로 SAOC 복호화기에 전송되는 $G_K[b]$과 $G_V[b]$ 파라미터들을 통해서 계산된다. 만약 임의의 주파수 밴드 b에서 신호 내의 보컬 성분의 마스킹 임계치 $T_{K_V}[b]$가 음악 성분의 파워 $\sigma_{K_M}^2[b]$보다 크다면 해당 주파수 대역에서는 보컬 신호에 의해 음악 신호가 마스킹 되어 들리지 않게 된다. 그러므로 방해 신호의 파워가 교차 예측을 수행한 뒤, 추정된 신호 y_D 내의 보컬 성분의 마스킹 임계치 $T_{y_{D^-}V}[b]$가 $\sigma_{K_M}^2[b]$같도록 설정해준다면 추정 알고리듬이 수행된 뒤에도 적어도 청각적으로 음악 성분에 대해서는 영향을 끼치지 않을 수 있다. 이 조건을

만족하는 $\sigma_D^2[b]$를 계산하는 과정은 식(19)와 식(20)에 나타내었다. 나아가, $\sigma_{K_V}^2[b]$가 $\sigma_{K_M}^2[b]$에 비해 두드러지게 높다면, $T_{K_V}[b]$와 $\sigma_{K_M}^2[b]$의 차이도 커진다. 즉, $T_{K_V}[b] \gg \sigma_{K_M}[b]$라고 생각할 수 있다. 이는 교차 예측을 통해 억제 시킬 수 있는 보컬 성분의 양도 많아지므로, 제안된 알고리듬의 효과를 극대화 시킬 수 있다.

$$
\begin{aligned}
&\text{for } b = 1 : B \\
&\quad \text{if}(T_{K\_V}[b] > \sigma_{K\_M}^2[b]) \\
&\qquad calc\ T_{D\_V}[b] = \sigma_{K\_M}^2[b] \\
&\quad end \\
&end \\
&T_{y_D\_V}[b] = \sigma_{K\_M}^2[b] \\
&\frac{\sigma_{y_D\_V}^2[b]}{O[b]} = \sigma_{K\_M}^2[b] \\
&\frac{G_V^2(1-|\rho[b]|^2 A[b])\sigma_K^2[b]}{O[b]} = G_M^2\sigma_K^2[b] \\
&A[b] = \frac{1}{|\rho[b]|^2}\left(1-(O[b]\frac{G_M^2[b]}{G_V^2[b]})\right), \\
&\qquad\qquad A[b] = \min(A[b], 1),
\end{aligned}
\tag{19}
$$

$$
\sigma_D^2[b] = \left(\frac{1}{A[b]} - 1\right)\sigma_S^2[b]. \tag{20}
$$

식(20)은 식(9)를 참조하여 식(19)에서 계산 된 $A[b]$값을 대입하여 구할 수 있다.

Fig. 4는 SAOC 복호화된 신호의 파워 스펙트럼을 임의의 프레임에 대해서 제안된 보컬 추정 알고리듬을 사용하기 전과 후의 결과를 나타낸 것이다. SAOC 복호화할 때 사용된 다운믹스는 1.5 kHz, 2 kHz의 톤(tone) 신호와 노이즈 신호로 구성되어 있으며, 양자화를 수행하지 않았다. 실험 과정에서 톤 신호와 노이즈 신호는 각각 독립된 보컬과 음악 신호로 간주되었다.

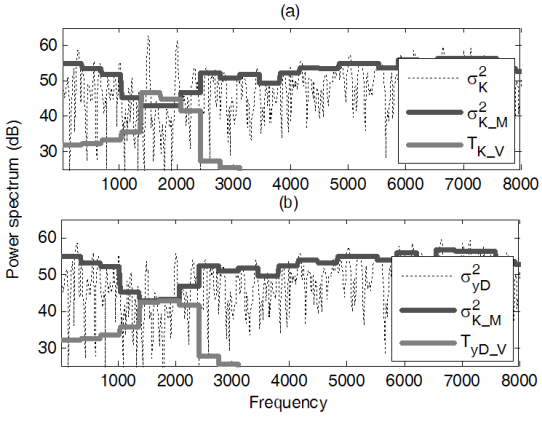

Fig. 4. Comparison of the power spectra of (a) the SAOC output in Karaoke mode (y_K), (b) the output of the vocal suppression filter (y_D).

Fig. 4에서 결과는 8 kHz까지만 나타내었다. 보컬 추정 알고리즘은 톤 신호가 남아 있는 밴드에 대해서만 적용되었으며, 잔여 톤 신호들은 식(19)를 통해 제거되었다. Fig. 4의 (a)를 보면 몇몇 밴드에서 $T_{K_V}[b]$가 $\sigma_{K_M}^2[b]$보다 높은데, 이는 해당 대역에 대해서 잔여 톤 신호의 파워가 상대적으로 커서 노이즈 신호가 마스킹 되었음을 의미한다. 이러한 대역에 대해서 제안된 알고리듬이 적용되어 보컬이 제거된 뒤, 출력 신호내의 보컬 성분의 마스킹 임계치 $T_{y_D_V}[b]$를 확인하면, 해당 대역에 대해서 음악 성분의 파워 크기만큼 낮아졌음을 Fig. 4(b)를 통해 확인할 수 있다. 이는 추정된 신호내의 보컬 성분이 줄어들었음을 의미한다. 그러므로 Fig. 4를 통해서 비록 SAOC 복호화된 신호내의 보컬 성분은 완전히 제거되지 못했지만, 각 대역에 대해서 청각적으로 음악의 음질을 열화시키지 않았음을 알 수 있다.

Fig. 5는 Fig. 4의 결과를 좀 더 자세히 관찰하기 위해, 보컬 추정 알고리듬이 수행되는 임의의 대역에 대해서 출력 신호의 스펙트럼을 관찰하였다. 그림에서 노이즈 신호 $y_{ref}[n]$를 Karaoke모드에서의 정답 신호로 간주하고, 톤과 노이즈 신호가 섞인 다운믹스

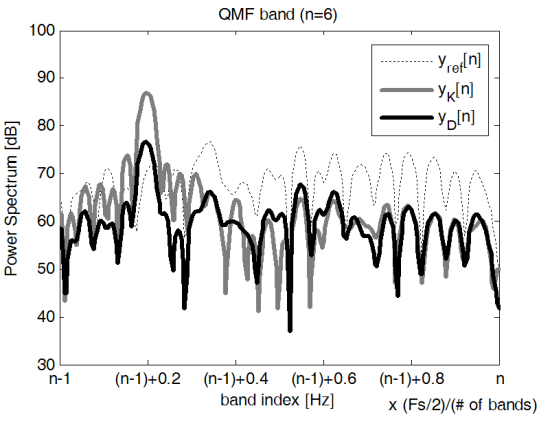

Fig. 5. Comparison of the power spectrum of the output at an arbitrary band.

신호로부터 복호화된 신호를 $y_K[n]$, 제안된 알고리듬까지 적용된 결과를 $y_D[n]$로 나타내었다. $y_D[n]$의 스펙트럼을 보면 $y_K[n]$에 남아 있는 톤 신호(x축에서 약 (n-1)+0.2에 위치함)가 보컬 추정 알고리듬을 통해 효과적으로 제거되었음을 알 수 있다. 나아가, $y_D[n]$ 스펙트럼을 보면 톤 성분이 추정된 위치로부터 멀어질수록 스펙트럼의 전체적인 변화가 작아져 톤이 위치한 곳에서만 음질의 변화가 발생하는 것을 확인할 수 있다.

IV. 실험 결과

제안된 알고리듬의 성능을 평가하기 위해, 객관적 및 주관적 음질 평가를 수행하였다. 제안된 알고리듬은 합성 QMF가 수행되기 전에 적용되었으며, 예측 계수는 4차로 설정하였다. 입력된 오디오 객체들은 부호화단을 통해 모노 신호로 다운믹스 하였다. SAOC 복호화를 진행할 때에는 양자화된 다운믹스 신호가 사용되지만, 제안된 알고리듬의 효과만을 확인하기 위해 양자화하지 않은 신호도 함께 실험에 포함하였다. 평가를 진행할 때에는 음악 객체만으로 구성된 $y_{ref}[n]$ 신호를 생성하여, Karaoke 신호 $y_K[n]$와 보컬 추정 알고리듬이 적용된 신호 $y_D[n]$를 비교하며 평가하였다.

실험에 사용된 SAOC의 부호화기는 MPEG에서 제공되는 Reference model[8]을 개량하여 사용하였으며, 양자화된 다운믹스 신호는 HE-AAC를 이용하여 20kbps의 전송률로 부호화하여 사용되었다. 실험 샘플들은 Table 1에 나타내었다. "TN"을 제외한 샘플들은 MPEG 표준화 회의 때 공식 실험용으로 제안되었던 샘플들이며, 모두 44.1 kHz의 샘플링 레이트를 갖는다.

Table 1의 "Rock" 샘플의 경우, 다른 샘플들과는 달리 모든 악기들이 같은 공간에서 동시에 연주된 샘플이다. 즉, 5.1 채널 신호가 6개의 객체로 이루어져 있다고 생각할 수 있다. "TN"은 톤과 노이즈 신호를 이용한 인위적으로 구성된 샘플이다.

객관적인 실험는 세그멘탈 SNR(segmental SNR)로 평가 되었다.

$$SEGSNR_X = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{\frac{1}{N}\sum_{n=0}^{N-1} y_{ref}^2(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}[y_{ref}(n+Nm)-y_X(n+Nm)]^2} \quad (21)$$

식에서 M과 N은 각각 프레임과 샘플 수를 의미한다. 아래 기입된 X는 출력 신호의 타입을 나타낸다. Table 2에 각 샘플에 대한 결과를 나타내었다.

표에서 Unquantized와 Quantized는 각각 다운믹스 신호의 양자화를 수행하지 않은 경우와 수행한 경우를 의미한다. 결과를 보면 SNR_{y_K}과 SNR_{y_D}의 차이는 크지 않는 것으로 나타났다. 하지만, 이는 전체 프레임에 대해서 평균을 낸 결과이며, 제안된 알고리듬은 각 주파수 밴드에 대해서 선택적으로 수행된다. Table 2의 가장 오른쪽 열에 임의의 주파수 밴드에 대해서 차이가 가장 컸던 결과를 함께 나타내었다. 실험 결과, 모든 신호에 대해서 대체로 큰 증가를 보이고 있는데, 이러한 증가는 제안된 알고리듬이 샘플 신호 곳곳에 적용된 구간들에 대해 음질 향상을 기대할 수 있는 결과로 생각될 수 있다.

SAOC는 제한된 전송률 및 파라메트릭 환경에서 부호화가 진행되어 복호화된 뒤에는 음질의 열화가 발생할 수 있다. SAOC의 Karaoke 모드는 여러 객체들 안에서 특정 객체를 추출하여 복원하는 것이므

Table 1. List of test items.

Item	Description (objects)
Hit	Bass, Guitar, Keyboard, Rhythm, Vocal
Sad promise	Bass, Brass, Chorus, Guitar, Rhythm, String, Vocal
K-pop01	Bass, Drum, Piano, Guitar, Organ, Vocal
Rock	L, R, C, LFE, Ls, Rs, Vocal
TN	Tone, Noise

Table 2. Segmental SNR.

Item	$SNRy_D$-$SNRy_K$ (Unquantized)	Max diff. in band (Unquantized)	$SNRy_D$-$SNRy_K$ (Quantized)	Max diff. in band (Quantized)
Hit	0.30	4.25	0.24	3.55
Sad promise	0.26	6.68	0.38	6.60
K-pop01	0.22	3.83	0.32	3.36
Rock	0.23	2.93	0.10	1.64
TN	1.96	4.68	1.67	4.11

로, 음원 분리(source separation) 분야에서 추구하는 목적과 비슷하다. 그러므로 SAOC에서도 음원 분리할 때와 마찬가지로 음질의 왜곡, 간섭 및 뮤지컬 노이즈 등이 발생 할 수 있다.[9] 그러므로 본 논문에서는 제안된 알고리듬의 성능 향상을 확인하기 위하여 위에서 언급된 열화 및 전체적인 음질에 대해서 주관적 음질평가를 수행하였다. 총 8명의 숙련된 실험자가 참여하였으며, 모두 헤드폰을 착용하고 실험을 진행하였다. 각 청취자들에게 음질의 열화 및 전체적인 음질은 ITU-R BS.1284 권고안[10]에 따라 Table 3에 표시되어 있는 지표를 참조하도록 하였고, 음악

Table 3. Quality and impairment scale.[10]

Grade score	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

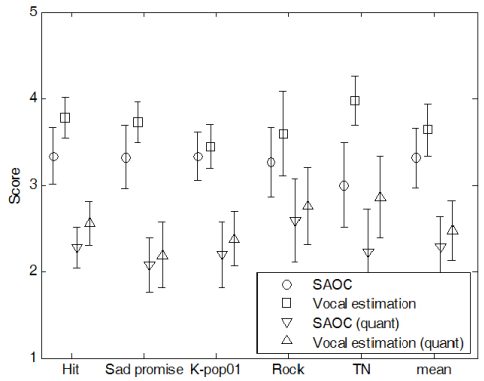

Fig. 6. Subjective evaluation results (overall quality).

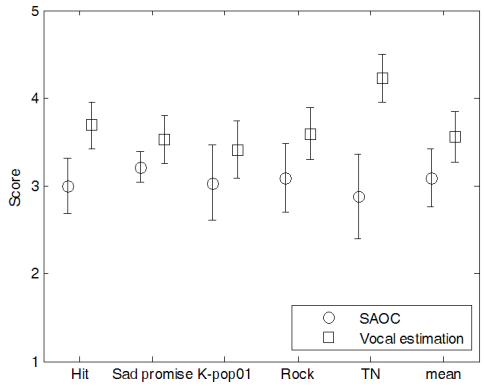

Fig. 7. Subjective evaluation results (interference).

객체만으로 구성된 $y_{ref}[n]$ 신호를 기준으로 하여, $y_K[n]$와 $y_D[n]$를 서로 비교하면서 평가하도록 하였다. 음질의 열화에 대해서 실험할 때에는 다른 부가적인 영향을 최소화하기 위해 양자화하지 않은 다운믹스 신호를 이용하였다. 실험 결과, 음악 음질의 왜곡 및 뮤지컬 노이즈의 경우 제안된 알고리듬을 적용하기 전과 후의 평가 결과가 크게 차이가 없어 영향을 덜 끼치는 요소들로 판단하여 본 논문에서는 제외하였고, 음질의 간섭 및 전체적인 음질에 대해서만 Fig. 6과 Fig. 7에 결과를 표시하였다.

실험 결과의 평균을 계산할 때에는 인위적으로 만든 실험 샘플 "TN"을 포함시키지 않았다. 실험 결과를 보면 제안된 알고리듬을 이용하면 복호화된 신호 내에 존재하는 보컬 성분을 효과적으로 억제시킬 수 있음을 확인할 수 있다. 실험 샘플 "HIT"의 경우, 음악 성분에 비해서 보컬 성분의 에너지가 절대적으로 높은 구간들이 존재하여, SAOC의 Karaoke 모드로 복호화하여도 보컬성분의 에너지가 큰 구간들이 존재하지만, 이러한 경우에도 제안된 알고리듬을 이용하면 보컬 신호를 효과적으로 억제시켜 눈에 띄는 성능 향상을 확인할 수 있다.

제안된 알고리듬을 적용하기 위해서는 복호화단에서 파라메트릭 연산을 통해 추가로 보컬 신호(SAOC의 Solo 모드) 및 교차 예측을 수행하여야하며, 이로 인해 연산량이 증가된다. 그러나 파라메트릭스 스테레오(parametric stereo) 부호화의 연산량을 고려해보면, 파라메트릭 연산은 QMF의 연산과 비교하여 충분히 낮은 것을 알 수 있고,[11] 교차 예측의 차수 역시 높지 않기 때문에 제안된 알고리듬으로 인해 SAOC 복호화의 전체적인 연산량이 크게 증가되지 않는다고 판단된다.

V. 결 론

본 논문에서는 SAOC의 Karaoke 모드에서 복원된 신호 내에 존재하는 보컬 성분을 추정하여 억제시킴으로써 전체적인 음질을 향상시키는 알고리듬을 제안하였다. 잔여 보컬은 SAOC의 Karaoke 모드로 복원된 신호와 솔로 모드로 복원된 신호와의 교차 예측을 통해 추정이 가능하지만, 두 신호사이의 상관성

이 매우 높아, 복원에 필요로 하는 음악성분까지도 함께 추정되어 제거될 수 있다. 이러한 열화를 방지하기 위해 보컬을 추정과정에서 예측을 억제하는 예측 방해 신호를 추가하였으며, 이 신호의 크기는 사람의 청각적인 마스킹 특성을 고려하여 음악적 음질의 열화가 최소화가 되도록 적응적으로 조절하였다. 객관적 및 주관적 음질 평가 결과 제안된 알고리듬은 Karaoke 신호 내에 남아있는 보컬 성분을 효과적으로 제거하여 음질을 향상시킬 수 있음을 확인하였다. 이러한 개선과정에서 제안된 알고리듬은 부호화단에서 어떤 추가 정보를 필요로 하지 않는 장점을 지닌다.

참 고 문 헌

1. ISO/IEC 23003-2:2010, *Information technology-MPEG audio technologies-Part 2: Spatial Audio Object Coding (SAOC)*, 2010.
2. G. Hotho, L. Villemoes, and J. Breebaart, "A backward-compatible multichannel audio codec," IEEE Trans. on Audio, Signal and Language. Proc., **16**, 83-93 (2008).
3. C. Falch, L. Terentiev, and J. Herre, "Spatial audio object coding with enhanced audio object separation," Proc. of the 13th Int. Conf. on DAFx-10 (2010).
4. J. Park, J. Hong, K. Kim, and M. Hahn, "Harmonic elimination structures for Karaoke mode in spatial audio object coding scheme," IEEE Int. Conf. on Consum. Elec.,813-814 (2011).
5. J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers and W. Oomen, "Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding," 124th AES Conv., paper no. 7377 (2008).
6. J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen and S. Van De Par, "Background, concept, and architecture for the recent MPEG surround standard on multichannel audio compression," J. Audio Eng. Soc., **55**, 331-351 (2007).
7. J. D. Johnston, "Transform coding of audio signal using perceptual noise criteria," IEEE Journal on Sel. Areas in Commun., **6**, 314-323 (1988).
8. ISO/IEC JTC1/SC29/WG11 N11037, *Study on ISO/IEC FCD 23003-2:200x, Spatial Audio Object Coding*, 2009.
9. V. Emiya, E. Vincent, and N. Harlander, V. Hohmann, "Multicriteria subjective and objective evaluation of audio source separation," 38th AES Int. Conf. : Sound Qual. Eval. (2010).
10. Recommendation ITU-R BS.1284-1, *General methods for the subjective assessment of sound quality*, 2003.
11. E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, "Low complexity parametric stereo coding," 124th AES Conv., paper no. 6073 (2004).

저자 약력

▸ 이 동 금 (Tung Chin Lee)

2005년 : 연세대학교 전기전자공학과 (학사)
2007년 : 연세대학교 전기전자공학과 (석사)
2008년~현재 : 연세대학교 전기전자공학과 (박사과정)
〈관심분야〉 디지털 신호처리, 오디오 신호처리, 음성 신호처리

▸ 박 영 철 (Young-cheol Park)

1986년 : 연세대학교 전기전자공학과 (학사)
1988년 : 연세대학교 전기전자공학과 (석사)
1993년 : 연세대학교 전기전자공학과 (박사)
2002년~현재 : 연세대학교 컴퓨터정보통신공학부 교수
〈관심분야〉 디지털 신호처리, 오디오 신호처리, 음성 신호처리, 적응 신호처리

▸ 윤 대 희 (Dae Hee Youn)

1977년 : 연세대학교 전자공학과 (학사)
1979년 : Kansas State University (석사)
1982년 : Kansas State University (박사)
1985년~현재 : 연세대학교 전기전자공학과 교수
〈관심분야〉 디지털 신호처리, 오디오 신호처리, 음성 신호처리, 적응 신호처리