컨벌루션 신경망을 이용한 공간큐 기반 다채널 오디오 확장 기술

백승권, 임우택, 이태진
한국전자통신연구원
{skbeack, wtlim, tjlee}@etri.re.kr

Seungkwon Beack, Wootaek Lim, Tajin Lee
Electronics and Telecommunications Research Institute (ETRI)

요 약

본 논문에서는 컨볼루션 신경망을 이용하여 추출 된 공간 오디오 큐를 이용한 오디오 채널 확장 기술을 소개한다. 오디오 채널 확장 기술은 일반적인 스테레오 신호에 적용되어 5.1 레이아웃과 같은 고차원 오디오 신호를 생성하는 기술이다. 스테레오 신호에서 채널을 확장하기 위해 스테레오 신호에서 공간 큐를 추출하고 추출 공간 큐의 정보에 따라 5.1 채널 신호의 스펙트럴 구성 요소를 할당하여 다중 채널 신호를 합성한다. 제안된 방식으로 생성된 5.1 채널 신호는 원 5.1 채널과 유사한 공간 정보 합성 능력과 스테레오 대비 주관적 선호도가 개선된 음질을 제공한다

1. 서론

본 연구에서는 음질의 열화 없이 일반 스테레오 오디오 콘텐츠로부터 5.1 채널 오디오 신호로 확장하는 기술을 소개한다. 다채널 오디오를 생성하는 기술은 오디오 채널확장기술에 기인하는데, 주로 주요한 정보를 추출하기 위한 개선 수단으로 활용되어 왔다 [1,2].

본 연구의 목적은 실제 확장된 5.1 채널 오디오 신호가 원 5.1 채널 콘텐츠와 같이 유사한 공간감과 몰입감을 제공할 수 있기를 희망한다. 이는 UHD 방송 서비스와 같이 영상 서비스가 콘텐츠의 차별화로부터 서비스의 차별화를 달성한 것처럼, 오디오 콘텐츠도 다채널 서비스를 보편적으로 활용할 수 있도록 하여, 다양한 오디오 서비스를 실현하고자 함이다. 최근 UHD 방송표준에는 MPEG-H 3D Audio 라는 최신 오디오 표준 기술이 반영되었으며, 다양한 포맷의 오디오 콘텐츠를 최상으로 표현할 수 있는 오디오 재현 기술이 포함되어 있다[3]. 따라서 오디오 콘텐츠 제작자의 관점에서 음상 정위에 무리가 없는 다채널 오디오 콘텐츠를 제공한다면 UHD 방송 서비스를 위한 차별화된 오디오 서비스를 제공받을 수 있을 것이다.

본 연구에서는 이러한 목적을 달성하기 위하여 실제 콘텐츠 제작자가 제공하는 다량의 5.1 콘텐츠를 수급하고, 이를 활용하여 스테레오 콘텐츠로부터 5.1 채널 콘텐츠를 생성하는 알고리즘을 연구하였다[4]. 5.1 채널로 확장하는데 중요한 정보는 5.1 채널이 스테레오 채널과 구별되게 가지고 있는 정보가 공간 정보이며, 이를 1 차적으로 예측하고, 예측된 공간 정보를 기반으로 5.1 채널을 생성하되 음질 열화를 최소화 할 수 있도록 하였다. 예측 되는 공간 정보는 공간큐(Spatial Audio Cue) 정보로 활용하여 기존 다채널 오디오 코딩 방식에 적용하여 다채널 오디오 신호를 생성하였다.

생성된 오디오 신호는 객관적, 주관적 평가를 통해 성능 검증을 수행하였으며, 최종적으로 공간 정보의 성공적인 예측과 주관적 선호도 음질에서 개선된 결과를 나타내었다.

2. 공간큐 기반 오디오 채널 합성 기술

오디오 채널의 공간 정보는 크게 세가지로 구성된다[5]. 채널간 레벨 정보(CLD: Channel Level Difference), 채널간 상관성(ICC: Inter-Channel Coherence), 그리고 채널간 시간 지연 (ILD: Inter-Channel Delay)이 주요 공간 정보이며 이를 공간큐라 정의한다. 모노 오디오 신호로부터 스테레오 신호를 예측할 경우, CLD, ICC, ILD 모두 활용되나, 공간상에 재생을 목적으로 하는 다채널 오디오 신호의 경우, 즉 스테레오 이상의 레이아웃 신호의 합성 시에는 ILD 는 활용되지 않는데, 이는 ILD 가 지각적으로 공간상에서 인지하기 쉽지 않기 때문이고, ICC 가 일부분 그 기능을 수행할 수 있기 때문이다[5,6,7]. 본 논문에서는 CLD 와 ICC 를 분석 및 예측하고자 한다. 기본적으로 스테레오 신호로부터 CLD 와 ICC 정보를 추출할 수 있으며, 이를 입력 데이터로 활용하여 미지의 다채널 오디오 신호의 CLD 와 ICC 값을 예측한다. CLD 와 ICC 는 다음과 같이 정의한다.

$$CLD_b^{ch} = 10\log_{10}\frac{P_{ch,b}}{P_{ch+1,b}} \tag{1}$$

$$ICC_b^{ch} = \frac{\mathrm{Re}\left(ch_b^f \odot ch_{b+1}^f\right)}{\sqrt{P_{ch,b}+P_{ch+1,b}}} \tag{2}$$

여기서, b 는 서브밴드에 대한 인덱스이며, ch 는 채널에 대한 인덱스이다. 서브밴드의 수는 48kHz 표본주파수를 고려할 때, 20 개 혹은 28 의 밴드로 구성되며, ch 는 합성하고자 하는 채널 pair 당 할당되는 채널 인덱스이다. 예를 들어, 5.1 채널을 합성하고자 한다면, 스테레오 신호로부터 4 개의 추가 채널 신호를 합성해야 하므로, $0 \le ch \le 4$ 이며, 서브밴드 수를 고려하여 추출 및 예측을 수행한다. 한 쌍의 공간큐 정보로부터 2 개의 채널 신호를 합성할 수 있는데, 합성하는 방식은 다음과 같다 [5,6].

$$\begin{bmatrix} ch \\ ch+1 \end{bmatrix} = \underbrace{\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}^t}_{\mathbf{H}} \begin{bmatrix} DMX \\ D(DMX) \end{bmatrix} \tag{3}$$

수식 (2)를 살펴보면, DMX 는 임의의 다운믹스 신호로, 일반적인 스테레오 신호의 좌/우 채널 신호로 간주할 수 있다. $D(DMX)$는 DMX 로부터 동일한 주파수 크기 성분을 가지면서, 위상 정보만 대체하여 DMX 와 상관성이 떨어지는 가상의 채널 신호를 생성하는 동작이다. 즉, DMX 입력신호로부터 DMX, $D(DMX)$를 생성하여 H 행렬을 적용하여 두개의 채널을

생성하는데, 하나의 입력으로부터 두개의 출력을 생성하는 동작으로 OTT(One-To-Two) 합성 모듈이라 정의한다. OTT 는 다채널 오디오 신호를 생성하는 근간이 되는 동작 모듈로 다수의 연결된 OTT 로부터 다양한 레이아웃을 갖는 다채널 오디오 신호를 생성할 수 있다. MPEG Surround 의 경우 5.1 채널을 스테레오 신호로부터 생성하기 위하여 4 개의 OTT 연산을 수행한다[10]. H는 CLD 와 ICC 파라미터로부터 구해질 수 있는 계수 및 위상 정보에 대한 행렬정보이다. 본 논문에서는 4 개의 CLD, ICC 파라미터 셋을 각 서브프레임마다 예측하여 이를 MPEG Surround 합성 과정에 적용하여 5.1 채널을 생성하였다.

3. 제안하는 공간큐 예측 신경망 구조

5.1 채널 신호를 생성하기 위하여 본 논문에서는 해당 채널의 공간큐를 예측하고 이를 이용하여 다채널 오디오 신호를 생성한다[7]. 다채널 오디오 신호를 생성하는 과정은 예측된 공간큐를 MPEG Surround 실제 파라미터로 간주하고 MPEG Surround 복호화기의 데이터로 활용하여 출력 신호를 생성하였다. 예측해야 하는 공간큐는 CLD 와 ICC 파라미터로 컨볼루션 신경망을 이용하여 예측한다. 컨볼루션 신경망 구축에 앞서 정의되어야 하는 것은 신경망의 입력 데이터 형상과 예측하고자 하는 대상에 대한 정의이다. 예측하고자 하는 대상은 공간큐 정보이므로, 신경망의 출력은 공간큐 정보를 예측한 값으로 표현되어야 할 것이다. 이를 위해서, 훈련데이터는 실제 5.1 채널 콘텐츠로부터 채널 pair 별 공간큐를 추출하고 각 채널의 공간큐 정보를 양자화 하여 신경망 출력 정보에 매칭 되어야할 레이블 정보로 활용하였다. 양자화 정보로 매칭한 이유는, 신경망을 통해 얻고자 하는 정보는 CLD 와 ICC 의 임의의 값이 되어야 하나, 임의의 실수 값을 예측하는 대신에, 각 공간큐의 양자화된 양자화 인덱스를 예측하는 것으로 대신함으로써, 예측 대상의 범위를 좁힐 수 있고, 뿐만 아니라 신경망의 학습을 분류기 형태로 학습할 수 있다는 장점이 있다. 분류기의 분류된 클래스 정보는 양자화 인덱스 값으로 간주할 수 있으며, 양자화 인덱스 값으로부터 해당되는 CLD, ICC 값으로 변환한다. 이러한 시도가 성능에 미치는 영향을 예상할 수 있는데, 이는 기존의 공간큐 기반 다채널 오디오 코딩 방식에서 활용되고 있는 부호화 전략을 그대로 따르고 있다고 볼 수 있다. MPEG Surround 등은 CLD 의 값을 31 개의 값으로 양자화 하며, ICC 값은 10 개의 값으로 양자화 한다. 마찬가지로, 본 논문에서 제안하는 신경망의 출력은 31 개의 CLD 값을 분류하기 위한 클래스와 ICC 값을 분류하기 위한 10 개의 클래스 값으로 출력 값을 얻을 수 있다.

공간큐를 예측하기 위한 신경망을 분류기 형태로 구축하였으나, 본 분류기의 출력 값은 상호간에 상관성이 존재하는 특이한 경우이다. 일반적인 분류기는 각 클래스 간에 상관성이 크게 존재하지 않는데, 실측 값을 예측하기 위한 분류기는 각 분류기 인덱스 값이 가까울수록 실측 값과 비교하여 적은 에러율을 나타낸다. 이는 출력된 클래스 인덱스가 양자화기 인덱스이기 때문이며, 양자화기가 스칼라(scalar) 양자화기로 각 클래스를 추정한 것이라면, 인덱스 간의 상관관계는 가까울수록 크다고 볼 수 있다. 이러한 특성을 손실함수에 반영하여 다음과 같이 정의하였다.

$$Loss(t)=\sum_{b}^{Nb}\left((1-\alpha)\cdot\sum_{j}^{Q}r_j(b)\log\frac{1}{p_j(b)}+\alpha\cdot\left(C_{r_j}(b)-C_{y_j}(b)\right)^2\right) \quad (4)$$

시간 t 에 대한 손실 함수는 두개의 항으로 구성되어 있다. 첫번째 항은 일반적인 클래스 문제를 풀기 위한 cross-entropy 손실 항이며[8], 두번째 항은 클래스 인덱스간의 거리 차를 나타내는 손실 항이다. $r_j(b)$ 는 b 번째 서브밴드의 j 번째 양자화 인덱스로, 양자화 인덱스를 one-hot 벡터로 표현한 요소값이다. $p_j(b)$ 는 해당하는 클래스의 확률값으로 출력된 값이다. $C_{r_j}(b)$ 는 실제 양자화 레벨 값을 정수형으로 표현한 값으로 예측된 양자화 레벨 값이다. $C_{y_j}(b)$ 는 해당 예측 값의 정답인 레이블 양자화 정수 값이다.
손실 함수의 구성과는 별도로 신경망의 구성은 자유롭게 구성할 수 있다. 단, 신경망의 출력이 분류기 형태로 구성되어 있으면 된다. 본 논문에서는 inception 구조의 컨볼루션 신경망을 채택하였다[12]. inception 구조를 채택한 이유는, 시간/주파수 축으로 다양한 형태의 컨볼루션 필터를 구성할 수 있으며 신호의 분석을 오디오 특성에 적합하게 구성할 수 있기 때문이다. 구축된 신경망 구조는 표 1 과 같이 나타내었다.

표 1. 공간큐 예측을 위한 컨볼루션 신경망 구조

층	입력형태	필터형태	출력형태
입력	[28,28,4]	[5,5]	[28,28,32]
은닉 (x4)	[28,29,32]	[1,1]x4	[28,28,32]x4
	[28,28,32]	[5,5],[5,1],[1,5]	[28,28,32]x3
	[28,28,128]	[3,1]	[14,14,128]
출력	[1,1,8192]	[1,1]	[1,1,868]

4. 실험 결과

4.1 데이터 베이스 및 실험 환경

제안된 신경망을 학습하기 위하여 원 5.1 콘텐츠를 DVD/Blu-ray 디스크를 통해 확보하였다. 300 개의 상이한 장르의 타이틀로부터 500 시간 이상의 원본 5.1 오디오 신호를 추출하였으며, 유효한 구간을 설정하여 주요 클립 3000 개를 추출하고 각 클립 별로 레이블 작업을 수행하였다. 각 클립의 구성은 30~60 초 가량의 재생 시간을 가지며, 이중 1000 개의 클립을 본 논문의 알고리즘 학습을 위해 활용되었다. 90:10 비율로 훈련데이터와 검증데이터로 활용하였으며, 최종 테스트 데이터는 MPEG 에서 제공하는 5.1 평가용 데이터를 활용하였다. 평가용 데이터는 총 100 개의 클립으로 구성되어 있으며, 각 클립 별로 20 초 내외의 길이를 갖는다.

4.2 음상 정위 정확도 평가

제안된 공간큐 예측 알고리즘의 성능을 평가하기 위해서 음상 정위 예측 정확도를 평가하였다. 음상정위 예측 정확도는 객관적 지표에 근거하여 측정되는 것으로, 예측된 파라미터로부터 정위되는 공간 정보 위상 값을 구하고 이를 원 데이터의 공간정보 위상 값과 비교한다. 예를 들어, 원 5.1 채널의 센터채널과 전방 좌측 채널 간은 30 도 간격을 가지며,

CLD 정보로부터 레벨 크기를 구하여 위상 값을 0 도에서 30 도 사이에 매핑시키고 이를 가상 음원 위치로 정의한다. 위상 값의 변환은 아래 수식과 같다[9].

$$\theta_b^{ch} = \tan^{-1}\frac{\hat{P}_{ch,b}}{\hat{P}_{ch+1,b}} \quad (5)$$

여기서 서브밴드 b 의 파워정보인 $\hat{P}_{ch,b}$ 는 예측된 CLD 값으로부터 복원된 값이다. 예측 위상 정보가 원 위상정보와 유사하다면 위상정보로부터 생성되는 가상 음상정위가 유사하다고 판단할 수 있다. 표 2 는 위상정보로부터 추정된 가상 음상 정위에 대한 평균 오차를 나타낸다. 전방 음상에 대해서는 3 도 내외의 오차를 나타내며, 후방 음상에 대해서는 4 도 정도의 오차를 나타낸다. 즉, 전방에 음원의 위치가 원음에서 표현되는 것과 비교할 때, 3~4 도의 오차범위내에서 스테레오로부터 가상 음원을 추정하여 다채널 신호를 생성할 수 있다.

4.3 음상 정위 정확도 평가

제안 알고리즘의 성능을 측정하기 위하여 주관적 음질 평가도 수행하였다. 주관적 음질 평가를 위해서 실제 합성된 5.1 채널 신호가 필요하다. 이를 위하여, MPEG Surround 디코더를 활용하였다[10]. MPEG Surround 디코더는 다운믹스 스테레오 신호와 예측된 공간 큐를 입력 받아 다채널 오디오신호를 생성한다. 생성된 5.1 오디오 신호는 원본과 음질 비교를 수행하여 성능평가를 수행하였다. 주관적 성능평가 수행방법으로 MUSHRA(MUltiple Stimuli with Hidden Reference and Anchor) 방식을 따랐다[11]. 청취 평가 테스트 아이템은 MPEG 테스트 아이템에서 11 개를 선정하여 테스트를 수행하였으며, 총 11 명의 피험자가 참여하였다. 비교 시스템으로는 5.1 채널 확장에 활용된 스테레오 다운믹스 신호와, 5.1 원본을 숨은 참조 모델 (hidden reference)로 활용하였으며, 3.5 kHz 대역폭 제한 5.1 신호를 앵커(anchor) 모델로 활용하였다. 평가결과는 표 3 과 같다. 본 연구에서 주안점을 두었던 음질 왜곡 문제를 비교하고자 원 스테레오 음원과 음질 및 음상을 비교한 결과 실제 스테레오 콘텐츠보다 음질이 열화 되었다고 느끼는 사례는 발생하지 않았다. 대신에 공간감의 증가로 원 5.1 신호 대비 70 점 이상의 효과를 나타내는 것으로 평가되었다.

표 2. 가상 음상정위 측정 범위 및 정확도

측정 범위	오차범위
전방음상정위	∠3.19 ±0.63
후방음상정위	∠4.18 ±0.57

표 3 주관적 음질 성능 평가 결과

95%신뢰구간 / 비교시스템	최저점	평균점	최고점
스테레오 시스템	54.9	57.1	59.1
제안 시스템	73.4	75.8	78.2

5. 결론

본 논문에서는 스테레오 콘텐츠를 이용하여 5.1 채널 오디오 신호를 생성하는 기술을 제안하였다. 제안 기술의 특징은, 다채널 오디오 정보를 추정하는데 있어서, 직접적인 신호를 예측하는 대신에, 공간 정보에 해당하는 공간큐를 예측하였고 예측된 공간큐를 기존 다채널 오디오 코덱을 활용하여 5.1 신호로 생성하였다. 생성된 다채널 오디오 신호는 음상위치 예측 성능을 객관적으로 평가하여 유효하게 음상정위를 수행하고 있음을 판단할 수 있었으며, 주관적 음질 평가를 통해 스테레오 콘텐츠 대비 공간감을 포함한 향상된 음질을 제공하고 있음을 검증할 수 있었다. 추후 진행되어야 할 연구로는, 신경망 구조의 간소화로 복잡도를 최적화 할 수 있어야 할 것이며, 실시간 처리에 대한 요구사항을 만족하기 위한 지연시간 최소화에 초점을 맞추어 알고리즘을 개선해야 할 것이다.

감사의 글

이 논문은 2017 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2017-0-00046, 데이터 지능형 분석기반 고수준 정보추출 원천기술 연구)

참고문헌

[1] D. Fitzgerald, A. Liutkus, Z. Rafii, B. Pardo, L. Daudet, “Harmonic/percussive separation using Kernel Additive Modelling”, Proc. of the 25th IET Irish Signals and Systems Conference, 2014.

[2] Wootaek Lim, and Taejin Lee. "Harmonic and percussive source separation using a convolutional auto encoder." 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017.

[3] J. Herre, et al. "MPEG-H 3D Audio-The New Standard for Coding of Immersive Spatial Audio." IEEE J. of selected topics in Signal Processing, vol. 9, No.5, Aug. 2015, pp. 770-779.

[4] S. Beack, W. Lim, T. Lee, "Spatial-Cue Based Audio Channel Extension Using Convolutional Neural Networks," Broadband Multimedia Systems and Broadcasting (BMSB), 2019 IEEE International Symposium on , June 2019

[5] J. Breebaart, et al. "MPEG spatial audio coding/MPEG surround: Overview and current status." Audio Engineering Society Convention 119. Audio Engineering Society, 2005, pp. 770-779

[6] S. Beack, et al. "MPEG Surround Extension Technique for MPEG-H 3D Audio." Etri Journal 38.5 (2016): 829-837.

[7] C. Faller, et al. "Binaural cue coding-Part II: Schemes and applications." IEEE Transactions on Speech and Audio Processing 11.6 (2003): 520-531.

[8] Bengio, Yoshua, et al. "Greedy layer-wise training of deep networks." Advances in neural information processing systems. 2007.

[9] K. Kim, et al. "Improved channel level difference quantization for spatial audio coding." ETRI journal 29.1 (2007): 99-102.

[10] ISO/IEC 23003-1:2007, MPEG-D (MPEG audio technologies), Part 1: MPEG Surround , 2007.

[11] International Telecommunication Union, “Method for the subjective assessment of intermediate sound quality (MUSHRA),” 2001, ITU-R, Recommendation BS, 1543-1, Geneva, Switzerland

[12] C. Szegedy, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." Thirty-First AAAI Conference on Artificial Intelligence. 2017.