JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 12, pp. 3811-3820, Dec. 2024

Check for updates

이미지 내 다중 음원 객체 분석을 통한 스테레오 오디오 생성

박 형 희[1] · 변 혜 원[2*]
[1]성신여자대학교 대학원 미래융합기술공학과 석사과정
[2]성신여자대학교 AI융합학부 교수

Image-to-Stereo Audio Generation from Multiple Audio-Source Objects

Hyung-Hee Park[1] · Hae-Won Byun[2*]

[1]Master's Course, Department of Convergence Technology Engineering, Sungshin Women's University, Seoul 02844, Korea
[2]Professor, School of AI Convergence, Sungshin Women's University, Seoul 02844, Korea

[요 약]

최근 인공지능 분야에서 이미지로부터 오디오를 생성하는 Image-to-Audio (I2A) 기술이 주목받고 있다. 그러나 기존의 I2A 기술은 단일 객체에 대응하는 단일 채널 오디오 생성에 중점을 두어 다중 객체가 포함된 이미지에서는 음원의 혼합 문제와 음원 누락 문제가 발생하는 한계를 보인다. 본 연구에서는 이러한 한계를 극복하고자 다중 객체를 포함하는 이미지에서 다중 음원 오디오를 스테레오로 생성하는 새로운 방법을 제시한다. 제안 모델은 YOLOv5를 활용하여 이미지 내 다중 음원 객체를 탐지하고, AudioLDM을 통해 각 객체에 대응하는 단일 음원을 생성한 후, 객체의 크기와 위치 정보를 기반으로 스테레오 오디오로 변환한다. 성능 평가를 위해 다중 객체 이미지-다중 음원 오디오 데이터셋을 새롭게 구축하였으며, 제안 모델은 다중 음원 객체 이미지에 대한 오디오 생성 시 모든 지표에서 베이스라인 모델보다 우수한 성능을 보였다. 이를 통해 기존 I2A 기술의 한계를 극복하고 다중 음원 객체를 포함한 복잡한 시나리오에서도 효과적으로 작동함을 입증하였다.

[Abstract]

Image-to-audio (I2A) technology has recently gained significant attention in the field of artificial intelligence. However, existing methods primarily focus on single-channel audio generation based on input images consisting of a single object, leading to issues such as blending of sound sources and missing audio elements in images consisting of multiple objecs. To address these limitations, we propose a novel approach for generating multi-source stereo audio from images containing multiple sound-producing objects. We employ YOLO (You Only Look Once) to detect multiple sound-producing objects and AudioLDM to generate distinct audio for each detected object. Subsequently, these individually generated audio sources are converted into stereo audio based on the sizes and positions of the corresponding objects. We evaluated the proposed model on a custom-built dataset comprising multiobject images paired with multisource audio, and it outperformed all baseline models across all metrics. This research overcomes the limitations of current I2A technologies by effectively handling complex, multisource scenarios, thereby advancing the field of audio generation.

색인어 : 이미지-오디오 생성, 스테레오 오디오, 다중 객체 음원 탐지, 멀티 모달 AI, 생성형 AI

Keyword : Image-to-Audio Generation, Stereo Audio, Multiobject Sound-Source Detection, Multimodal, Generative AI

http://dx.doi.org/10.9728/dcs.2024.25.12.3811

Received 31 August 2024; **Revised** 07 October 2024
Accepted 14 October 2024

***Corresponding Author; Hae-Won Byun**

Tel: +82-2-920-7615
E-mail: hyewon@sungshin.ac.kr

Ⅰ. 서 론

게임, 온라인 교육, AR/VR/XR 등과 같은 뉴 미디어 콘텐츠와 메타버스 플랫폼에서는 시각적 요소와 함께 음향적 요소가 사용자 경험에 중요한 영향을 미친다. 사실적이고 입체감 있는 음향은 시각적 장면과 결합하고 동기화되어 사용자의 이해도와 몰입도를 향상하는 데 도움을 준다. 이러한 시청각 미디어 경험에 대한 수요 증가로 인해 이미지로부터 오디오를 생성하는 Generative AI 연구에 대한 필요성이 대두되고 있다.

기존의 Image-to-Audio[1]-[3] 연구는 단일 객체 이미지로부터 해당 객체에 대응하는 한 개의 음원(Sound Source)을 생성하는 데 집중했다. 이 접근법은 단일 객체를 다루는 경우에는 유용하지만, 다수의 객체를 포함하는 이미지를 처리할 때 여러 한계를 보인다. 첫째, 다중 음원의 부적절한 혼합으로 인한 음원 혼합 문제가 발생할 수 있다. 둘째, 주요 객체의 음원이 생성되지 않는 음원 누락 문제가 발생할 수 있다. 또한, 기존 연구에서는 단일 음원을 단일 채널 오디오로 생성하여 복잡한 장면 내 객체들의 크기와 공간 내에서의 위치 정보를 적절히 반영하지 못해 생성된 오디오가 단조롭고 현실감이 떨어진다.

본 연구에서는 기존 연구를 확장하여 다중 객체들을 포함하는 이미지로부터 다중 오디오를 생성하는 새로운 방법을 제안한다. 이미지 내 소리가 나는 객체(음원 객체)에 대응하는 적합한 음원을 생성하고, 생성된 다중 음원을 스테레오 오디오로 조화롭게 통합한다. 왼쪽과 오른쪽 2개의 채널을 갖는 스테레오 오디오를 통해 이미지에 적합한 공간감과 입체감을 표현한다. 그림 1은 기존 연구의 접근 방식을, 그림 2는 본 연구에서 제안하는 방법을 비교하여 보여준다. 제안 모델은 다중 객체(예: 강아지와 새)를 포함하는 이미지를 입력으로 받아, 각 객체에 해당하는 개별 음원을 생성하고 2채널 스테레오 오디오를 출력한다.

본 연구에서 제안하는 시스템은 이미지에서의 음원 객체 탐지, 단일 객체의 음원 생성, 객체의 위치 및 크기 기반 다중 음원 통합의 세 부분으로 구성된다. 음원 객체 탐지 부분에서는 YOLO(You Only Look Once)[4] 모델 중 YOLOv5을 사용하여 객체를 탐지하고, 이후 탐지된 객체 중에서 음원 객체를 선별한다. 단일 객체의 음원 생성 부분에서는 AudioLDM[5]을 활용하여 각 음원 객체별 음원을 생성한다. 다중 음원 통합 부분에서는 각 객체의 크기와 위치 정보를 이용하여 오디오의 좌우 볼륨을 조절하고 객체별 음원을 통합하여 스테레오 오디오를 생성한다.

본 연구에서 기여하는 바는 다음과 같다.

첫째, 단일 객체를 포함하는 이미지로부터 단일 음원을 생성하는 기존의 Image-to-Audio 방법을 확장하여 다중 객체를 포함하는 이미지에서 다중 음원을 생성하는 방법을 제안하였다.

둘째, 각 객체의 크기와 위치를 기반으로 음원의 볼륨을 조절하여 스테레오 오디오를 생성함으로써 청취자의 몰입도를 향상시킬 수 있다.

셋째, 다중 객체 이미지와 이에 대응하는 다중 음원 오디오로 구성되는 데이터셋을 구축하고 공개하여 관련 분야의 연구를 활성화하는데 기여하였다.

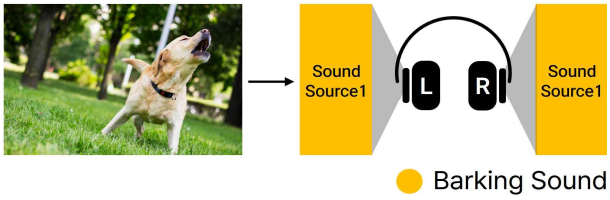

그림 1. 기존 Image-to-Auido Generation의 입출력
Fig. 1. Existing Image-to-Audio Generation models inputs and outputs

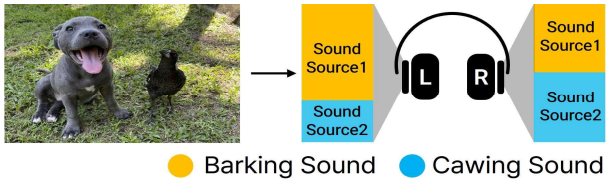

그림 2. 제안 모델의 입출력
Fig. 2. Proposed model inputs and outputs

Ⅱ. 관련 연구

2-1 Object Detection & Classification

객체 탐지(Object Detection)는 이미지에서 여러 객체를 찾아내고 해당 객체들의 위치를 나타내는 경계 상자(Bounding Box)를 출력하는 문제이며, 객체 분류(Object Classification)는 이미지나 영상에서 특정 위치에 있는 객체가 무엇인지 분류하는 문제이다. 최근에는 객체 탐지와 객체 분류를 함께 처리하는 딥러닝 모델들이 다양하게 제시되고 있다.

Faster R-CNN[6]은 RPN(Region Proposal Network)을 사용하여 입력 이미지에서 객체가 있을 가능성이 높은 영역을 빠르게 제안하고, 이후 CNN을 사용해 제안된 영역을 분류하여 객체의 종류를 결정한다. 이 모델은 RPN이 객체가 있을 가능성이 높은 영역을 정밀하게 제안하므로 정확성이 높지만, 단계적인 접근 방식과 복잡한 계산과정으로 비교적 속도가 느려 실시간 객체 탐지에는 적합하지 않다.

YOLO는 이미지를 그리드로 나누고, 각 그리드 셀이 하나의 객체를 포함한다고 가정하여, 객체의 탐지와 분류를 하나의 네트워크에서 동시에 수행하여 매우 빠른 속도로 결과를 도출한다.

SSD(Single Shot MultiBox Detector)[7]는 YOLO처럼 단일 단계에서 객체 탐지를 수행하면서, 이미지 그리드 셀마다 다양한 크기와 비율의 사각형을 사용하여 다양한 크기의

객체를 동시에 예측함으로써 속도와 정확성의 균형을 이루고자 하였다.

본 연구에서는 YOLOv5을 활용하여 이미지 내 다중 객체의 경계 상자(Bounding Box)의 좌표를 예측했다. YOLOv5는 적은 연산 자원을 사용하면서 동시에 객체 탐지의 정확성을 유지하므로 오디오 생성의 전체 처리 시간을 단축하는 데 효과적이다. 또한, 쉬운 구현 방법은 연구 개발 과정에서 효율성을 높여준다.

2-2 Image-to-Audio Generation

Image-to-Audio Generation[1]-[3]은 최근 생성형 AI 연구에서 주목받고 있는 분야로 이미지를 입력으로 오디오를 생성하는 기술이다.

SpecVQGAN[3]은 이미지에서 오디오를 생성하는 과정에서 중간 표현으로서 이미지의 특징을 반영하는 오디오 스펙트로그램을 사용한다. 스펙트로그램을 효과적으로 생성하기 위하여 벡터 양자화(VQ)를 도입하여 이미지 데이터를 잠재 벡터로 압축 표현하였고, 스펙트로그램에서 오디오로의 변환에서 오디오 생성의 품질을 높이기 위해 GAN(Generative Adversarial Network)을 적절하게 결합하였다.

Im2Wav[2]은 스펙트로그램을 사용하지 않고 이미지에서 오디오 신호를 직접 생성하며, 이미지와의 의미적 일관성을 유지하면서 고품질 오디오를 생성하는 데 중점을 두었다. 사전 학습된 CLIP(Contrastive Language-Image Pre-training)[8]의 이미지 인코더를 활용하여 이미지에서 관련 특징을 추출하고, 이를 Transformer 모델의 입력으로 사용하였다. 2개의 Transformer를 도입하여 저수준의 오디오 신호를 먼저 생성하고, 이를 고해상도 오디오로 업샘플링 하여 고품질의 최종 오디오를 생성했다.

최근에는 학습 부담을 완화하고, 대규모 학습을 통해 얻은 일반화 능력을 상속받기 위해 CLIP, CLAP(Contrastive Language-Audio Pre-training)[9], AudioLDM과 같은 기반 모델(foundation model)을 활용한 경량 솔루션이 제안되었다. 해당 연구에서는 CLIP의 시각적 특징을 CLAP의 오디오 특징으로 변환하는 맵퍼를 제안하여 도메인 간의 차이를 줄이고, 맵퍼를 통해 변환된 CLIP 임베딩을 사전 훈련된 오디오 생성 모델인 AudioLDM의 입력으로 사용해 적은 학습 비용으로 고품질의 오디오를 생성할 수 있음을 입증했다.

본 논문은 기존의 연구 방법들을 활용하여 이미지 내 여러 객체를 기반으로 다중 음원 오디오를 생성하는 새로운 접근 방식을 제안한다.

2-3 Stereo Audio Generation

스테레오 오디오는 2개의 채널을 사용해 소리의 공간적 배치를 재현하는 음향 기술이다. 사용자는 소리가 특정 위치에서 발생하는 것처럼 느낄 수 있으며, 이는 몰입감 있는 음향 경험을 제공하는 데 중요한 역할을 한다. 최근 시각 정보를 활용해 스테레오 오디오를 생성하는 연구가 진행되고 있다.

Sep-Stereo[10]는 스테레오 오디오 생성과 오디오 음원 분리가 유사한 목표를 가지고 있다는 점에 착안하여 두 작업을 단일 모델로 통합해 스테레오 오디오 생성 성능을 향상했다. PseudoBinaural[11]은 시각-스테레오로 구성되는 데이터셋을 제작하는 어려움을 완화하기 위해 단일 오디오를 사용하여 의사(pseudo) 시각-스테레오 쌍을 구성하고, 이를 활용해 실제 바이노럴 녹음 없이 바이노럴 오디오 변환 네트워크를 훈련시키는 방법을 제안했다. L2BNet[12]은 데이터 수집의 한계를 극복하기 위해 음원 위치 추정을 활용했다. 두 가지 네트워크를 사용하여 음원 위치 추정을 통해 약한 감독 학습을 수행함으로써, 학습에 필요한 바이노럴 오디오 데이터의 양을 줄였다. SAGM[13]은 생성적 적대 신경망(GAN)을 활용하여 시각 정보를 바탕으로 단일 오디오를 바이노럴 스테레오 오디오로 변환하는 방식을 제안했다. 비디오에서 추출한 시공간적 시각 정보와 단일 오디오의 특징을 결합하여 소리의 방향성과 깊이감을 재현하고, 생성된 오디오와 실제 오디오를 구분하는 판별기를 통해 생성기의 성능을 향상시켰다.

본 논문에서는 이미지 내 객체의 크기와 위치 정보를 활용해 단일 오디오를 스테레오 오디오로 변환하였다. 각 객체의 위치와 크기를 기반으로 왼쪽과 오른쪽 채널의 오디오 볼륨을 조절함으로써, 이미지 내 공간 정보를 효과적으로 반영한 스테레오 오디오를 생성했다.

Ⅲ. 연구 방법

본 연구에서 제안하는 모델 구조는 그림 3과 같다. 음원 객체 탐지, 음원 객체별 음원 생성, 객체 크기 및 위치 기반 스테레오 오디오 생성으로 구성된다.

3-1 음원 객체 탐지

객체는 이미지 내에서 시각적으로 식별할 수 있는 사물이며, 음원 객체는 오디오 생성에 활용될 수 있는 객체를 의미한다. 이미지 내 다중 음원 객체를 탐지하기 위해 객체 탐지 모델 YOLOv5를 도입하였다.

YOLOv5의 객체 탐지 결과, 그림 4(a)와 같이 음원 객체인 '기타'가 '기타' 자체의 독립 객체로 탐지될 뿐만 아니라 동시에 '사람' 객체의 일부로도 탐지되는 음원 객체 중복 탐지 현상이 확인됐다. 이와 같은 현상은 오디오 생성에서 불필요한 음원을 중복하여 생성하는 문제를 초래할 수 있다. 각 음원 객체에 대해 하나의 고유한 음원을 생성하기 위해 아래와 같은 방법으로 중복 객체를 탐지하고 처리했다.

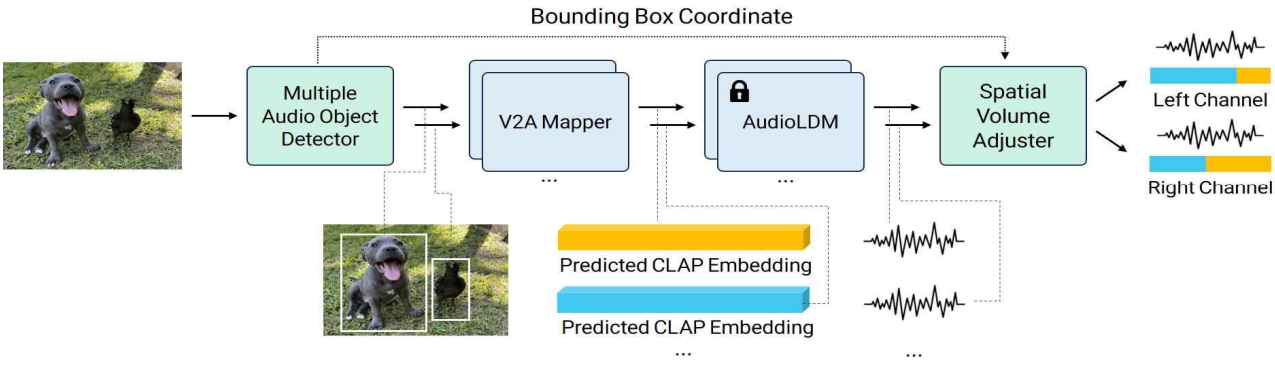

그림 3. 다중 음원 객체 이미지 기반 다중 음원 오디오 생성 시스템 구조도
Fig. 3. System architecture of a multi-sound source audio generation pipeline based on multi-object images

그림 4. 객체 탐지 결과: (a) 일반 객체 탐지, (b) 음원 객체 탐지
Fig. 4. Object detection results: (a) object detection, (b) sound source object detection

1) 중복 탐지 객체 검출

YOLOv5로 탐지한 모든 객체를 순회하여 다른 객체와 겹친 영역의 크기를 계산하였다. 다른 객체와 겹친 영역이 전체 영역의 50% 이상을 차지하는 경우, 중복으로 탐지되었다고 판단하였다.

2) 중복 탐지 객체 처리

중복으로 탐지된 객체 중 전체 넓이가 더 넓은 객체를 대표 객체로 선별했다. 최종적으로 선별된 대표 객체들은 해당 이미지의 음원 객체로 오디오 생성에 활용했다. 이 과정을 통해 중복 객체를 효과적으로 처리하여 오디오 생성 과정에서 발생할 수 있는 불필요한 중복 음원을 방지하였다.

3-2 음원 객체별 음원 생성

이미지로부터 오디오를 생성하기 위해 오디오 생성 모델인 AudioLDM과 V2A-Mapper를 활용하여 음원 객체별 단일 음원을 생성하였다.

AudioLDM은 Text-to-Audio Generation 모델로 CLAP의 텍스트 인코더로 입력 텍스트의 임베딩을 추출하고, 이를 활용해 오디오를 생성한다. CLAP은 대규모 텍스트-오디오 데이터셋을 기반으로 학습되어 텍스트와 오디오 간의 의미적 관계를 효과적으로 표현하여 오디오 생성의 조건으로 활용하기 적합하다. 그러나 이미지 기반 오디오 생성을 위해서는 이미지 특징을 CLAP의 오디오 임베딩 공간으로 변환하는 추가적인 단계가 필요하다.

이를 위해 V2A-Mapper를 도입하여 CLIP의 이미지 특징 벡터를 CLAP의 오디오 임베딩 공간으로 매핑하였다. 구체적으로, CLIP의 이미지 인코더를 사용하여 객체 이미지로부터 이미지 특징 벡터를 추출한 뒤, V2A-Mapper를 통해 이를 CLAP의 오디오 임베딩 공간으로 매핑한다. 매핑된 벡터를 AudioLDM의 입력으로 사용하여 오디오를 생성한다.

V2A-Mapper는 Multi-Layer Perceptrons (MLPs)로 구현하였으며, VGGSound[14] 데이터셋으로 학습했다. 그림 5는 VGGSound 데이터셋 중 5,000개의 데이터를 무작위로 샘플링한 뒤, CLIP과 CLAP 임베딩을 추출해 2차원 공간에 투영한 결과이다. V2A-Mapper 적용 전, CLIP과 CLAP 임베딩 간의 평균 유사도가 -0.0073으로 매우 낮았던 것과 달리 V2A-Mapper 적용 후 코사인 유사도가 0.479로 크게 상승하였다. 이는 V2A-Mapper가 이미지와 오디오 도메인 간의 차이를 효과적으로 줄였음을 시사한다. 이를 통해 AudioLDM의 추가적인 학습 없이 고품질의 음원 객체별 개별적인 단일 음원 오디오를 생성하였다.

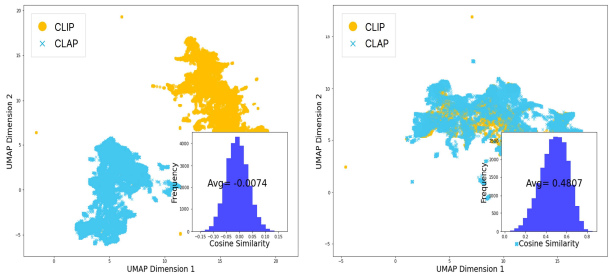

그림 5. CLIP 임베딩과 CLAP 임베딩 간 도메인 간극 변화: (a) V2A-Mapper 도입 전, (b) V2A-Mapper 도입 후

Fig. 5. Domain gap in CLIP image embeddings and CLAP audio embeddings: (a) before V2A-Mapper implementation, (b) after V2A-Mapper implementation

3-3 객체 크기 및 위치 기반 스테레오 오디오 생성

이번 절에서는 음원 객체별로 생성된 단일 채널 오디오를 스테레오 오디오로 변환하는 방법을 설명한다. 탐지된 음원 객체의 크기와 위치 정보를 활용하여 좌우 오디오 세기와 오디오 세기와 볼륨을 조절하여 스테레오 오디오를 생성하였다.

1) 객체 크기 기반 오디오 세기 조절

음원 객체의 크기 정보를 활용하여 음원의 좌우 소리 세기를 조절하였다. 이는 소리 세기 v는 객체와 관찰자 사이의 거리 d의 제곱에 반비례한다는 물리적 원리에 기반한다. 관찰자로부터 객체까지의 거리 d를 추론하기 위해 2차원 이미지에서 깊이 정보를 감지하는 단안 단서를 활용하였다. 큰 물체일수록 더 가깝다고 인식하는 '상대적 크기' 단서를 활용하여 객체의 크기 s와 거리 d 사이의 역비례 관계를 식 (1)과 같이 정리하였다. 객체의 크기가 큰 경우 해당 음원 객체에서 발생한 소리의 세기가 증가하고, 작은 경우 소리의 세기가 감소한다.

$$v \propto \frac{1}{d^2} \tag{1}$$

$$d \propto \frac{1}{\sqrt{s}}$$

$$\therefore v \propto s$$

2) 객체 위치 기반 오디오 볼륨 조절

음원 객체가 이미지 중심 x 좌표를 기준으로 어느 위치에 있는지에 따라 왼쪽과 오른쪽 오디오의 볼륨을 조절하였다. 음원 객체의 중심 좌표 vx는 -1과 1 사이의 값을 가지며, -1에 가까울수록 객체가 이미지 내의 왼쪽에, 1에 가까울수록 이미지 내 오른쪽에 위치함을 의미한다. vx의 절댓값이 1에 가까울수록 왼쪽과 오른쪽의 오디오 볼륨 차이가 크게 발생하며, 왼쪽 볼륨 vx_left와 오른쪽 볼륨 vx_right는 식 (2)으로 표현할 수 있다.

$$vx_{\text{left}} = 0.5 \times (1 - \text{vx}) \tag{2}$$

$$vx_{\text{right}} = 0.5 \times (1 + \text{vx})$$

최종 볼륨 파라미터는 v_size와 vx_left, vx_right를 조합하여 계산한다. 각 볼륨 파라미터의 가중치 w1과 w2는 실험을 통해 조정이 가능하나, 일반적으로 0.5로 설정하여 사용한다. 최종 볼륨 파라미터는 식 (3)과 같이 계산된다.

$$vx_{\text{left}} = \text{w}_1 \times \text{v}_{\text{size}} + \text{w}_2 \times \text{vx}_{\text{left}} \tag{3}$$

$$vx_{\text{right}} = \text{w}_1 \times \text{v}_{\text{size}} + \text{w}_2 \times \text{vx}_{\text{right}}$$

그림 6은 단일 음원 오디오에 볼륨 조절 알고리즘을 적용한 결과로, 최종 볼륨 파라미터를 단일 음원 오디오에 적용하여 음원 객체의 공간 정보를 청각적으로 표현할 수 있음을 보여준다.

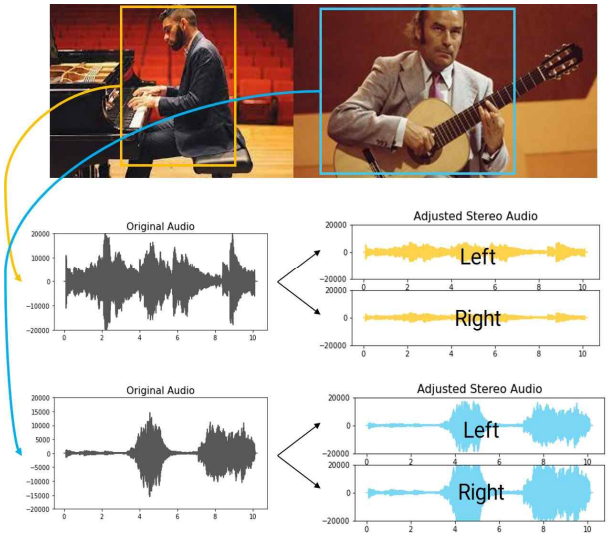

그림 6. 객체 크기 및 위치 기반 볼륨 조절 알고리즘 적용 결과

Fig. 6. Results of object size and position-based volume adjustment algorithm

Ⅳ. 실 험

4-1 실험 설정

제안 모델의 구현 및 실험에 사용된 주요 파라미터와 설정은 다음과 같다. 이미지 내 객체 검출을 위해 YOLOv5의 사전 훈련된 모델을 사용하였으며, IoU 임곗값은 0.45, Confidence 임곗값은 0.25로 설정하였다.

V2A-Mapper는 Linear Layer, SiLU 활성화 함수, 레이어 정규화를 순차적으로 적용하는 구조로 설계하였으며, 중간 레이어의 깊이 D와 확장률 E를 각각 4로 설정하였다. 학습은

1000 에폭 동안 진행하였으며, 옵티마이저는 AdamW을 사용하였고 학습률은 1.1e-4로 설정하였다. 멀티모달 처리를 위해 이미지 인코더로 CLIP의 "ViT-B/32" 버전을 채택하였고, 오디오 인코딩 및 생성에는 AudioLDM의 사전 훈련 모델을 활용하였다.

AudioLDM은 audioldm-s-full-v2 버전을 사용하여 10초 길이의 고품질 오디오를 생성하였다. 오디오 생성의 품질과 다양성을 최적화하기 위해 Guidance Scale을 2.5로 설정하였으며, DDIM 샘플링에는 200단계를 적용하였다. 추가로 전반적인 오디오 생성 품질 향상을 위해 n_candidate_gen_per_text 파라미터를 3으로 설정하여 각 객체에 대해 3개의 오디오 후보를 생성하고, 이미지와의 의미적 일치도가 가장 높은 오디오를 최종 선택하도록 하였다.

4-2 데이터셋

제안 모델의 평가를 위해 이미지와 다중 음원 스테레오 오디오 쌍으로 구성된 데이터셋을 구축하였다. 기존에 공개된 데이터는 주로 악기 소리에 국한되어 있어, 다양한 음원을 포함하는 데이터셋 구축을 목표로 하였다. 이를 위해 기존의 단일 이미지-단일 음원 데이터셋인 ImageHear[2]와 VGGSound[6]를 기반으로 [다중 음원 객체 이미지-다중 음원]의 쌍으로 구성된 Multi-VGGSound와 Multi-ImageHear 데이터셋을 새롭게 구축하고 공개하였다.

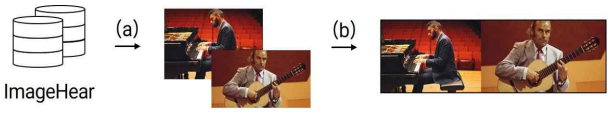

그림 7. Multi-ImageHear 데이터셋 구축 과정
Fig. 7. Creation process of the Multi-ImageHear dataset

1) Multi-ImageHear 데이터셋

30개의 클래스, 101개의 이미지로 구성된 ImageHear를 기반으로 Multi-ImageHear 데이터셋을 구축하였다. 그림 7은 Multi-ImageHear 데이터셋의 구축 과정을 보여준다. ImageHear에서 2~3개의 이미지 샘플을 무작위로 선택한 뒤, 선택 순서에 따라 횡 방향으로 병합하는 방식으로 120개의 다중 음원 객체 이미지를 생성했다.

2) Multi-VGGSound 데이터셋

Multi-VGGSound 데이터셋은 약 20만 개의 10초 길이 YouTube 비디오 클립으로 구성된 VGGSound를 기반으로 구축되었다. 그림 8은 Multi-VGGSound 데이터셋의 구축 과정을 보여준다.

데이터셋 구축에 앞서, VGGSound의 오디오 특성을 분석하기 위해 오디오 볼륨 분포를 살펴본 결과, 평균 -23.677, 분산 86.392로 오디오의 볼륨이 다양하게 분포되어 있음을 확인하였다. 여러 음원을 혼합해 다중 음원 오디오를 만드는 과정에서 특정 음원이 너무 크거나 혹은 작게 들리는 문제를 방지하기 위해 ffmpeg의 loudnorm 필터를 활용해 전체 오디오의 볼륨을 평균 -20.27, 분산 45.825로 정규화했다. 동영상 프레임은 1초에 1회 캡처하여 처음/중간/마지막 총 3장의 프레임을 추출하여 활용하였다.

VGGSound에서 2~3개의 샘플을 무작위로 선택한 뒤, 다중 음원 객체 이미지와 오디오를 생성했다. 이미지는 Multi-ImageHear와 동일하게 처리했으며, 오디오는 선택 샘플 수에 따라 다르게 생성했다. 2개의 샘플이 선택된 경우, 오디오를 순서에 따라 왼쪽과 오른쪽 채널에 삽입했다. 이때, 가운데 공간에 소리가 비는 것을 방지하기 위해 반대편 채널의 오디오 볼륨을 0.3배로 조절하여 중첩하였다. 샘플이 3개 선택된 경우, 두 번째로 선택된 샘플의 오디오 볼륨을 반으로 줄여 좌우 채널에 중첩하였다.

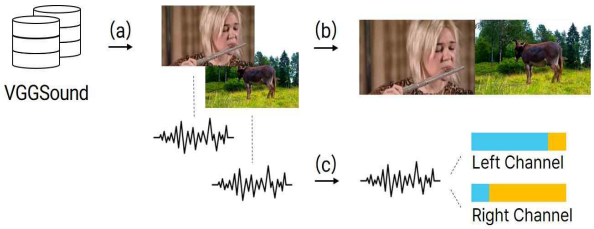

그림 8. Multi-VGGSound 데이터셋 구축 과정
Fig. 8. Creation process of the Multi-VGGSound dataset

4-3 평가지표

본 연구에서는 제안 모델의 성능을 정량적으로 측정하는 객관적 평가와 인간의 청각적 경험을 반영하기 위한 주관적 평가를 진행하여 다각도로 평가하였다.

1) 객관적 평가

객관적 성능 비교를 위해 Frechet Audio Distance(FAD), KL Divergence(KL), Clip-Score(CS)[2] 3가지 평가 지표를 사용했다.

FAD는 오디오의 전반적인 품질과 다양성을 평가한다. VGGish[15]를 임베딩 추출기로 활용해 생성 오디오와 원본 오디오의 임베딩 간 분포 거리를 계산한다. KL은 생성된 오디오와 원본 오디오의 확률 분포 간의 차이를 계산하여, 생성된 오디오와 실제 데이터의 유사성을 평가한다. 이때, Multi-VGGSound는 스테레오 오디오로 구성되어 있어 FAD와 KL 계산 시 각 채널에 대해 개별적으로 측정한 후 평균내어 사용하였다. CS는 입력 이미지와 생성 오디오 간의 의미적 관련성을 정량적으로 측정한다. CLIP의 이미지 인코더와 Wav2CLIP[16]의 오디오 인코더를 사용하여 이미지와 오디오 각각의 임베딩을 추출한 후, 이들 간의 코사인 유사도를 계산한다.

2) 주관적 평가

생성 오디오의 품질(Quality), 이미지와의 관련성(Relevance), 몰입도(Immersive) 3가지 측면에서 주관적 성능 평가를 진행했다. 각 데이터셋에서 20개의 이미지를 무작위로 선택하고 이를 조건으로 오디오를 생성한 뒤, 10명의 참가자에게 생성 오디오에 관한 평가를 요청했다. 참가자들은 각각의 오디오 품질에 대해서는 이산적인 5점 척도로 평가했으며, 관련성과 몰입도 측면에서는 참가자들의 더욱 명확한 의사 표현을 유도하기 위해 4점 척도로 평가하였다.

평가 결과는 각 모델의 평균 의견 점수(Mean Opinion Score, MOS)[18]로 요약되었으며, 그림 9는 주관적 평가에 사용한 평가지 일부를 캡처한 것이다.

V. 실험 결과

본 연구에서는 제안 모델의 성능을 평가하기 위해 기존 벤치마크 및 새로 구축한 데이터셋을 활용하여 SpecVQGAN, V2A-Mapper와의 비교 실험을 수행하였다. 이를 통해 제안된 방법론의 효과성을 검증하고자 하였다.

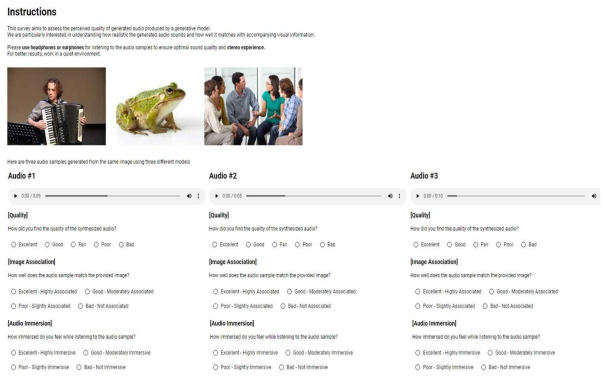

그림 9. Multi-ImageHear 데이터셋의 주관적 평가 설문지
Fig. 9. Subjective evaluation questionnaire for the Multi-ImageHear dataset

5-1 객관적 평가

표 1의 객관적 성능 평가 결과, 제안 모델은 단일 객체(VGGSound, ImageHear) 이미지에 대해 V2A-Mapper보다 낮은 성능을 보였다. 이는 음원 객체 탐지 과정의 오류가 생성된 오디오의 품질 저하에 유의미한 영향을 미친 것으로 판단된다.

그림 10은 탐지 오류 중 부적절한 객체 탐지의 두 가지 사례를 제시한다. 그림 10(a)은 중복 음원 객체가 제대로 처리되지 못한 사례로, 이 경우 동일한 음원이 중복 생성되어 최종 오디오에 비자연스러운 반향 효과를 초래할 수 있다. 그림 10(b)은 소리가 나지 않는 무음 객체가 탐지된 사례이다. 그림 11은 무음 객체 오디오의 스펙트로그램으로, 일부 소리가 발생한 것을 확인할 수 있다. 이러한 불필요한 오디오 생성은 최종 오디오에 노이즈로 작용하여 최종 오디오 품질의 전반적인 저하를 야기할 수 있다. 이와 달리, 그림 12는 중요 음원 객체가 탐지되지 않은 경우를 보여준다. 이 경우, 해당 이미지에 필수적인 음향 요소가 누락되어 최종 오디오의 품질과 완성도를 떨어뜨리는 요소로 작용하는 것으로 생각된다.

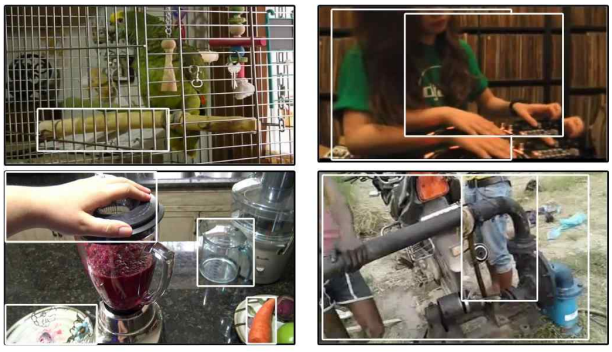

그림 10. 부적절한 음원 객체 탐지 사례: (a) 음원 객체 중복, (b) 무음 객체
Fig. 10. Error Case in Sound Object Detection: (a) Duplicate Sound Source Object, (b) Silent Object

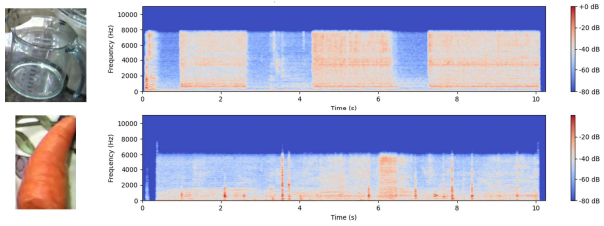

그림 11. 무음 객체로 생성된 오디오의 스펙트로그램
Fig. 11. Spectrogram of audio generated from silent objects

그림 12. 음원 객체 탐지 누락
Fig. 12. Missing sound source object detection

반면, Multi-VGGSound에 대해서는 제안 모델이 FAD, KL 지표에서 가장 좋은 성능을 보였다. 이는 다중 음원 객체 이미지에 대한 오디오 생성 시 제안 모델이 효과적인 방법임을 입증한다. CS 지표는 모든 데이터셋에서 V2A-Mapper가 가장 우수한 성능을 보였는데, 중요한 점은 Multi-VGGSound 데이터셋에 대한 성능 평가 결과에서 V2A-Mapper가 생성

표 1. 4가지 데이터셋을 활용한 최신 모델(SOTA)과의 객관적 성능 비교
Table 1. Objective comparison with SOTA methods on four datasets

Datasets / Models	VGGSound		Multi-VGGSound		ImageHear		Multi-ImageHear	
	FAD↓	KL↓	CS↑	FAD↓	KL↓	CS↑	CS↑	CS↑
Real Audio	0.0	−0.0004	9.737	0.0	−0.0005	7.468	–	–
SpecVQGAN	7.733	3.366	5.446	8.320	3.695	5.403	6.131	4.245
V2A-Mapper	3.593	2.592	9.113	4.021	3.079	8.660	9.228	8.086
Ours	3.915	2.991	8.048	3.451	2.873	6.865	4.386	7.923

표 2. 4가지 데이터셋을 활용한 최신 모델(SOTA)과의 주관적 성능 비교
Table 2. Subjective comparison with SOTA mMethods on four datasets

Datasets / Models	Multi-VGGSound			Multi-ImageHear		
	Quality↑	Relevance↑	Immersive↑	Quality↑	Relevance↑	Immersive↑
Real Audio	3.689	2.878	2.929	–	–	–
SpecVQGAN	2.407	1.929	1.888	2.467	1.862	1.892
V2A-Mapper	2.598	2.087	2.046	2.646	2.032	2.022
Ours	2.986	2.512	2.397	3.580	2.989	2.944

한 오디오의 CS(8.660)가 실제 오디오의 CS(7.468)보다 더 높게 측정된 것이다. 이러한 결과는 CS 계산에 사용된 이미지 및 오디오 인코더들이 단일 객체 및 단일 음원에 최적화되어, 다중 객체 및 음원의 특징을 효과적으로 포착하지 못한 데서 기인한 것으로 판단된다. 이로 인해 제안 모델의 CS 값이 전반적으로 과소평가 되었을 가능성이 존재한다.

5-2 주관적 평가

표 2에 제시된 바와 같이, 제안 모델은 Quality, Relevance, Immersive 세 가지 측면에서 모두 기존 모델을 상회하는 우수한 성능을 보였다. 이는 제안 모델이 다중 음원 객체 이미지에 대해 효과적으로 작동함을 시사한다.

주목할 만한 점은 객관적 평가에서 생성 오디오와 이미지 간 관련성(CS)이 낮게 측정된 것과 달리 주관적 평가에서는 제안 모델의 Relevance가 높게 평가되었다는 점이다. 이러한 평가 결과의 불일치는 객관적 평가와 주관적 평가를 병행하는 것의 중요성을 보여주며, 다중 음원 객체 이미지에 대한 오디오 생성 모델의 성능을 정확하게 평가하기 위한 새로운 평가 방법 개발의 필요성을 보여준다.

Ⅵ. 결 론

본 연구에서는 다중 음원 객체가 포함된 이미지로부터 다중 음원으로 구성된 스테레오 오디오를 생성하는 모델을 제안했다. 제안 모델은 YOLOv5가 탐지한 객체 중 일부를 음원 객체로 선별하여 단일 음원을 생성하였다. 이후, 음원 객체의 크기와 위치 정보를 통해 단일 음원을 스테레오 오디오로 변환한 뒤, 모든 음원을 결합하여 다중 음원 스테레오 오디오를 생성하였다. 실험 결과, 다중 객체 이미지에 대해 제안된 모델이 기존 모델 대비 우수한 성능을 보이며, 효과적으로 작동함을 확인하였다.

그러나 본 연구에서 제안한 방법은 몇 가지 한계점을 가지고 있다. 첫째, YOLOv5의 객체 탐지 능력에 의존함으로써 음원 객체 후보가 제한되는 문제가 있다. 특히 천둥번개나 비와 같이 시각적 객체로 명확히 정의되기 어려운 음원에 대한 오디오 생성에 한계가 있다. 둘째, 음원 객체 선별 과정에서 객체의 크기만을 고려하고 음향적 중요성을 반영하지 못하는 문제가 있다. 이로 인해 크기는 작지만 음향적으로 중요한 객체(예: 악기)가 더 큰 객체에 가려져 무시될 수 있다. 셋째, 음원의 볼륨 조절 과정에서는 객체의 깊이 정보를 고려하지 않아 부적절한 볼륨 설정이 이루어질 수 있다. 예를 들어, 원거리의 큰 객체가 근거리의 작은 객체보다 항상 더 큰 소리로 표현되는 현상이 발생한다.

이는 복잡한 3차원 장면을 청각적으로 표현하는 데 있어 중요한 제약 요소로 작용하며, 생성 오디오의 전반적인 품질을 저하시킬 수 있다.

향후 연구에서는 이러한 한계를 극복하기 위해 이미지 내 소리 발생 위치를 예측하고 이를 바탕으로 음원 객체를 탐지하는 방법을 제안하여 기존의 탐지 오류를 줄이고 전체 품질을 개선하고자 한다. 특히, 기존에 탐지가 어려웠던 음원 객체를 식별하고, 전경과 배경을 효과적으로 구분하여 배경 음원 생성함으로써 더욱 풍부하고 자연스러운 음향 환경을 구현하고자 한다. 추가로 모델의 파라미터 최적화를 통해 탐지 성능과 정확도를 향상시킬 계획이다. 또한, DINOv2와 같은 이미지 인식 모델을 활용하여 객체의 깊이 정보를 추출하고, 이를 음원 통합에 활용하여 기존의 오디오 세기 조절 기법의 한계를 보완하고자 한다. 마지막으로 현실적인 다중 음원 객체 데이터셋을 구축과 다중 객체 및 음원의 특징을 효과적으로 포착하는 평가 지표를 제안하여 다중 음원 오디오 생성 연구에

기여하고자 한다.

본 연구는 최신 생성형 AI 기술을 활용하여 기존 오디오 생성 기술의 한계를 극복하고, 다중 음원 오디오 생성 분야의 발전에 기여하였다는 점에서 의의가 있다.

감사의 글

이 논문은 2023년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.

참고문헌

[1] H. Wang, J. Ma, S. Pascual, R. Cartwright, and W. Cai, “V2A-Mapper: A Lightweight Solution for Vision-to-Audio Generation by Connecting Foundation Models,” in *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI-24)*, Vancouver, Canada, pp. 15492-15501, February 2024. https://doi.org/10.1609/aaai.v38i14.29475

[2] R. Sheffer and Y. Adi, “I Hear Your True Colors: Image Guided Audio Generation,” in *Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, pp. 1-5, June 2023. https://doi.org/10.1109/ICASSP49357.2023.10096023

[3] V. Iashin and E. Rahtu, “Taming Visually Guided Sound Generation,” arXiv:2110.08791, October 2021. https://doi.org/10.48550/arXiv.2110.08791

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-time Object Detection,” in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas: NV, pp. 779-788, June 2016. https://doi.org/10.1109/CVPR.2016.91

[5] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, ... and M. D. Plumbley, “Audioldm: Text-to-Audio Generation with Latent Diffusion Models,” in *Proceedings of the 40th International Conference on Machine Learning (ICML ’23)*, Honolulu: HI, pp. 21450-21474, July 2023. https://doi.org/10.48550/arXiv.2301.12503

[6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, June 2017. https://doi.org/10.1109/TPAMI.2016.2577031

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single Shot MultiBox Detector,” in *Proceedings of the 14th European Conference on Computer Vision (ECCV 2016)*, Amsterdam, Netherlands, pp. 21-37, October 2016. https://doi.org/10.1007/978-3-319-46448-0_2

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, ... and I. Sutskever, “Learning Transferable Visual Models from Natural Language Supervision,” in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, Online, pp. 8748-8763, July 2021. https://doi.org/10.48550/arXiv.2103.00020

[9] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, “Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-caption Augmentation,” in *Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, pp. 1-5, June 2023. https://doi.org/10.1109/ICASSP49357.2023.10095969

[10] H. Zhou, X. Xu, D. Lin, X. Wang, and Z. Liu, “Sep-Stereo: Visually Guided Stereophonic Audio Generation by Associating Source Separation,” in *Proceedings of the 16th European Conference on Computer Vision (ECCV 2020)*, Glasgow, UK, pp. 52-69, August 2020. https://doi.org/10.1007/978-3-030-58610-2_4

[11] X. Xu, H. Zhou, Z. Liu, B. Dai, X. Wang, and D. Lin, “Visually Informed Binaural Audio Generation Without Binaural Audios,” in *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville: TN, pp. 15480-15489, June 2021. https://doi.org/10.1109/CVPR46437.2021.01523

[12] K. K. Rachavarapu, A. Aakanksha, V. Sundaresha, and A. N. Rajagopalan, “Localize to Binauralize: Audio Spatialization from Visual Sound Source Localization,” in *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, pp. 1910-1919, October 2021. https://doi.org/10.1109/ICCV48922.2021.00194

[13] Z. Li, B. Zhao, and Y. Yuan, “Cross-Modal Generative Model for Visual-Guided Binaural Stereo Generation,” *Knowledge-Based Systems*, Vol. 296, 111814, July 2024. https://doi.org/10.1016/j.knosys.2024.111814

[14] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, “Vggsound: A Large-Scale Audio-Visual Dataset,” in *Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 721-725, May 2020. https://doi.org/10.1109/ICASSP40776.2020.9053174

[15] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke,

A. Jansen, R. C. Moore, ... and K. Wilson, “CNN Architectures for Large-scale Audio Classification,” in *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans: LA, pp. 131-135, March 2017. https://doi.org/10.1109/ICASSP.2017.7952132

[16] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, “Wav2CLIP: Learning Robust Audio Representations from CLIP,” in *Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 4563-4567, May 2022. https://doi.org/10.1109/ICASSP43922.2022.9747669

박형희(Hyung-Hee Park)

2023년 : 성신여자대학교 정보시스템공학과(공학사)

2023년～현 재: 성신여자대학교 대학원 미래융합기술공학과 석사과정

※관심분야 : 딥러닝, 멀티 모달, 생성형 AI

변혜원 (Hae-Won Byun)

2004년 : KAIST 대학원 (공학박사-컴퓨터그래픽스)

1992년 : KAIST 대학원 (공학석사)

1990년 : 연세대학교 전산과학과 (공학사)

2006년～현 재: 성신여자대학교 AI융합학부 교수

※관심분야 : 컴퓨터 그래픽스, 멀티모달 딥러닝, 생성형 AI