# #2 사운드 생성 AI가 불러올 미래

글. **전상배** 가우디오랩 CSO

### ChatGPT가 불러온 생성 AI 돌풍

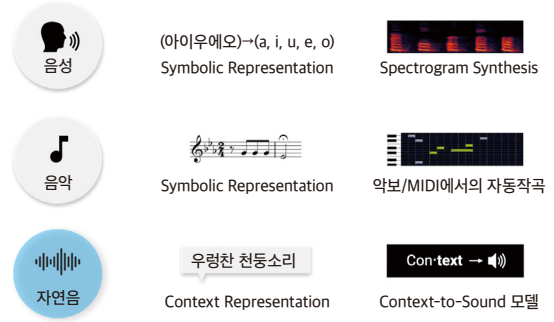
ChatGPT만큼 사람들의 삶에 빠르게 스며든 기술이 또 있었을까 싶을 정도로 생성 AI에 대한 관심이 뜨겁다. 마치 인터넷과 스마트폰 등장의 초창기를 보는 듯, 새로운 패러다임이 찾아오는 것이 아닐까 하는 기대까지 하게 되는 시점으로, ChatGPT에서 시작된 생성 AI 열풍이 글쓰기가 아닌 미술, 음악과 같은 다른 콘텐츠 영역까지 생성형(Generatvie) AI의 영향력을 확장하고 있다. 가우디오랩은 그중에서도 생성형 AI를 활용하여 소리를 생성하는 기술을 성공적으로 개발해 사업화를 추진 중이다.

### Sound Generation과 자연음 생성의 어려움

우선, 소리의 영역에 대해서 정의하고자 한다. 소리는 물리적인 진동이 공기를 매개체로 전파되는 것이다. 사람에게는 이 진동에 의해 청각 기관과 신경망이 자극되는데, 이 물리적인 현상은 자연 발생적으로 존재하는 충격음, 마찰음이나 폭발음 등의 자연음, 동물과 사람의 발성 기관에 의한 울음소리와 언어에 해당하는 발성음, 나아가 사람의 인위적인 설계를 통하여 화성, 리듬, 멜로디로 구성한 음악까지 포함한다.

이러한 다양한 소리 중에서도, 자연음이 아닌 ▲사람에 의해 생성되는 발성음 ▲음악의 문자나 악보처럼 소리에 직접적으로 대응되는 소리 등은 기호적 표현(Symbolic Representation) 체계가 잘 갖추어져 있다. 즉, 문자나 악보에

**그림 1.** 소리별 Sound Generation의 특징



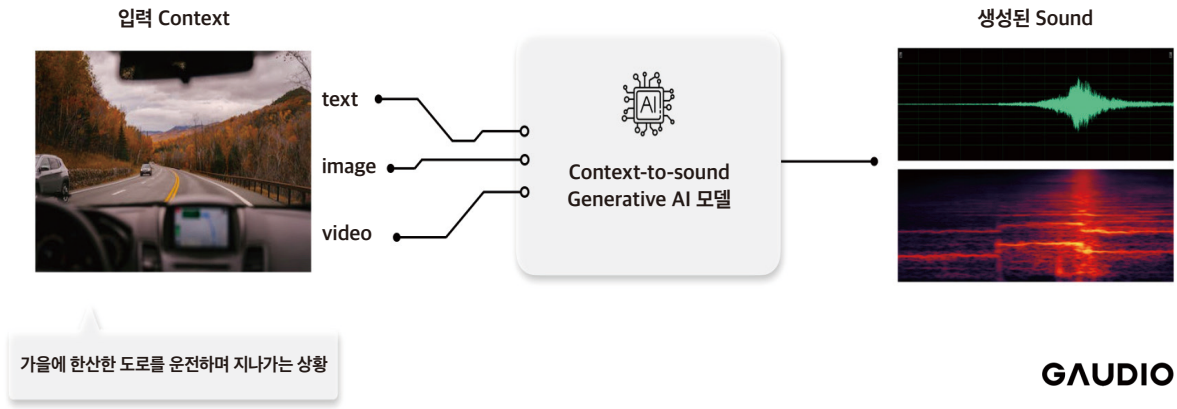

GAUDIO

소리와 직접적으로 상관도를 가지는 충분히 많은 정보가 표현되어 있고, 이를 활용하여 소리를 생성하는 것들은 Text-to-Speech(TTS)와 MIDI-based Music Generation과 같은 형태로 오랜 기간 동안 샘플링, 신호 처리와 AI 기술들을 활용하여 연구되어 왔다.

반면, 우리 주변에서 들려오는 무수히 많은 자연음에 대한 생성은 오랜 기간 동안 연구되지 않았다. 특정 소리의 생성 과정에 대한 물리적 모델링(Physical Modeling)으로 소리를 생성하는 연구는 있었지만, 세밀한 분석 작업을 통해 소리 발생 요소 하나하나를 모델링하는 물리적 모델(Physical Model) 기법의 특징상 현실적으로 존재하는 다양한 소리를 개별적으로 모두 모델링하여 생성하는 것은 매우 어려운 일이었다.

즉, 음성이나 음악과 달리 자연음은 발생 과정에서 매우 다

그림 2. 다양한 형태의 입력에 대응되는 Context 기반 AI Sound Generation




양한 매질과 현상에 기인하여 복합적으로 생성되는데, 이를 하나의 통합된 기호적 표현으로 구성하기는 사실상 불가능하다. 그래서 가우디오랩은 Context를 이해하고 이에 대한 소리를 생성하는, '불가능을 가능케 하는' 기술에 집중하게 되었다.

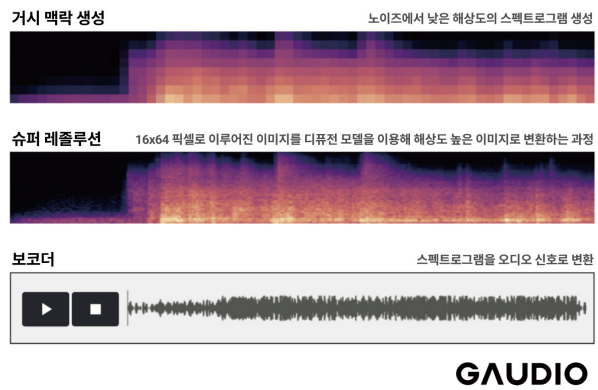
### 자연음을 생성하는 가우디오랩 AI Sound Generation

최근 생성형 AI의 활발한 연구 속에서, 다양한 소리에 대한 텍스트 묘사(Text Description)의 Context를 이해하고 이에 대응되는 소리를 생성하는 연구들이 출현하고 있다. 물론 가우디오랩도 자연음에 대응되는 Foley(효과음 녹음) 및 효과음을 생성하는 연구를 수행하고 있는데, 텍스트, 이미지, 비디오 상에서 정의하는 Context에 대응되는 소리를 생성하는 기술을 개발하고 있다. 현재 완성 단계에 있는, 텍스트로부터 소리를 생성하는 과정은 다음과 같다.

#### 1단계: 텍스트 입력에서 작은 스펙트로그램 생성하기

첫 단계에서는 만들고 싶은 소리에 대한 텍스트 입력을 처리한다. '우렁찬 천둥소리'라는 텍스트를 입력으로 받으면, 텍스트의 Context에 대응되는 임베딩으로 변환한 후, 해당 임베딩 정보를 활용하여 랜덤한 노이즈에서 작은 사이즈의 스펙트로그램을 만들어낸다. 이때 만들어진 스펙트로그램은 16×64 픽셀로 이루어진 작은 이미지로, 16개의 주파수 밴드와 64개의 프레임에 대한 소리를 의미하는 것으로, 어느 시점에 어떠한 크기 어느 음색을 가지며 소리를 내는가에 대한 거시적인 시간적 주파수적 맥락을 나타낸다.

그림 3. 가우디오랩 AI Sound Generation 구조



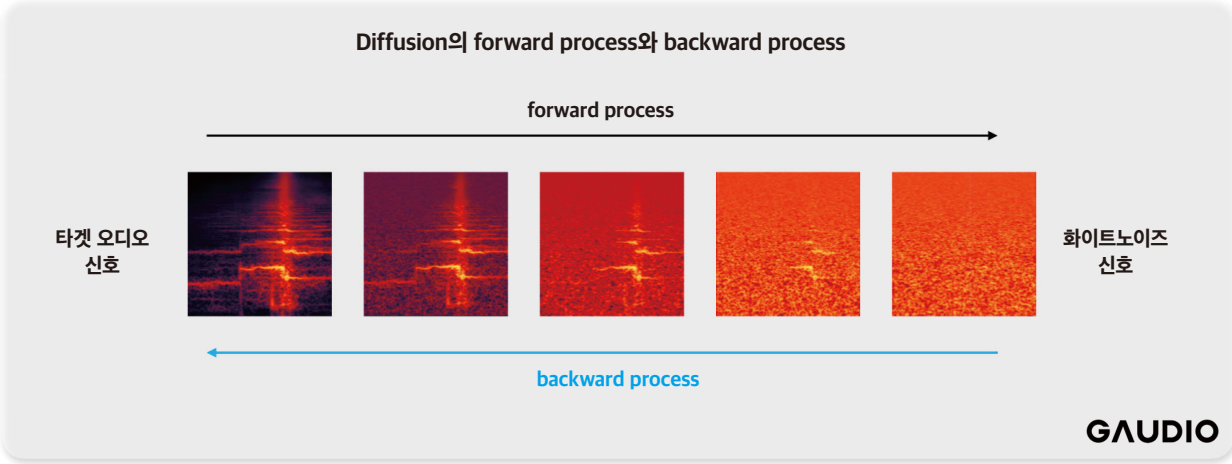

#### 2단계: 슈퍼 레졸루션(Super Resolution)

그 이후에는 이런 작은 이미지의 스펙트로그램을 점차 개선해 나가는 '슈퍼 레졸루션' 단계를 거쳐서 시간적, 음색적 완성도를 높여 나아간다. 이 과정에서도 1단계에서 활용된 임베딩도 같이 활용되어 해당 Context에 대응되는 디테일을 살린다. 그 결과 명료한 형태의 스펙트로그램이 완성된다.

#### 3단계: 보코더 (Vocoder)

이제 마지막 단계로 보코더를 활용하여 스펙트로그램을 오디오 신호로 변환해 준다. 통상적으로 보코더라 하면 음성 스펙트로그램으로부터 음성 신호를 만들어주는 기술을 의미하나, 자연음의 경우 소리에 대한 패턴이 음성과 다르기 때문에 이에 대한 보코더도 자체적으로 개발하였다.

그림 4. 생성 모델의 핵심인 Diffusion 개념




### 가우디오랩 AI Sound Generation의 핵심은 Diffusion

앞에서 설명한 가우디오랩의 AI Sound Generation 3단계에서 모두 Diffusion 기법을 사용하고 있다. Diffusion은 생성형 AI에서 널리 쓰이는 모델이다. 어떠한 신호가 소량의 백색소음(White Gaussian Noise)으로 오염되었을 때, 오염시키는 프로세스를 forward process, 정제되는 프로세스를 backward process라 정의하고, 조건부 확률 모델을 활용하여 backward process를 학습하여 신호를 생성해 내는 기법이다.

이러한 Diffusion 과정을 충분히 많이 반복하면 backward process를 통하여 백색소음으로부터 원하는 타깃 신호를 생성해 낼 수 있는데, 생성 과정에서 Context에 대응되는 임베딩을 Condition으로 활용하면 Context에 대응되는 방향으로 타깃 신호가 생성되는 원리이다.

따라서 소리를 생성하는 중요 조건(Condition)인 이 임베딩을 Text로부터 생성하면 Text-to-Sound, 이미지로부터 생성하면 Image-to-Sound, 영상으로부터 생성하면 Video-to-Sound가 되어 모든 정형/비정형 데이터로부터의 소리 생성이 가능해진다.

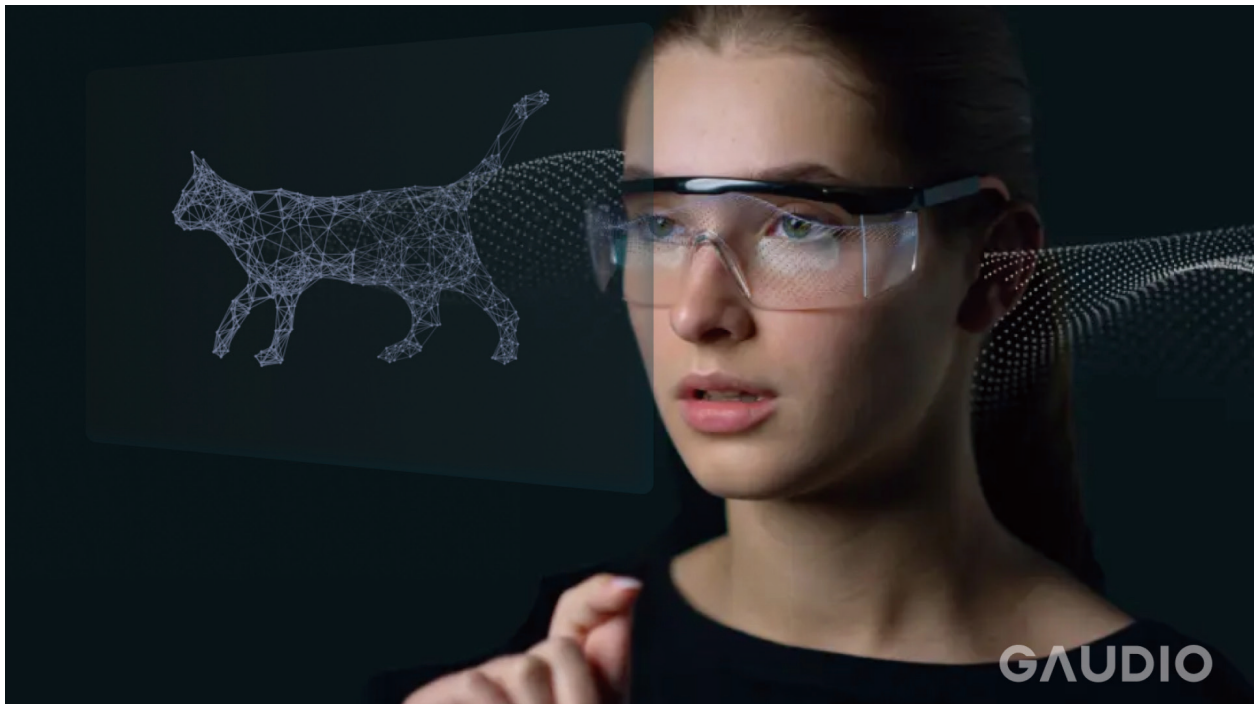
### 가우디오랩의 AI Sound Generation을 사용하면 무엇이 가능할까

가우디오랩의 AI Sound Generation은 텍스트, 이미지, 동영상의 다양한 도메인의 데이터로부터 그 의미 및 개념에 대응되는 Context를 유추하고, AI가 이에 해당하는 사운드를 생성하는 기술이다. 현재 주류 미디어인 영화, OTT 및 게임에서의 사운드 생성은 물론, 미래 콘텐츠 미디어인 메타버스에서의 사운드 제작 생산성을 혁신적으로 향상시킬 것으로 시장의 기대를 한 몸에 받고 있다. 또한, AI Sound Generation 자체뿐만 아니라 학습 과정에서 확보된 소리와 Context 간의 매핑은 기 제작 효과음들을 Context에 기반한 Library 화에 활용될 수 있다. 덕분에 리소스 관리 및 향상된 검색 기능을 제공할 수 있어 전반적인 사운드 작업의 생산성을 크게 향상시킬 것으로 기대한다.

#### 영화, OTT, 게임 등 주류 미디어 산업에서의 생산성 향상

일반적인 영화/OTT 콘텐츠의 사운드 후반 작업(Sound Post Production)은, 후시 대사, Foley 생성, 효과음 생성, 믹싱/마스터링의 과정을 거쳐서 진행된다. 대부분이 전문 사운드 엔지니어의 기억과 상식에 의존한 작업 프로세스로, 통상적으로 편당 4주 이상의 시간이 소요되기 때문에 규모가 큰 사운드 스튜디오의 경우에도 1년에 제작 가능한 양이 10편 내외에 머무른다. 2019년도 코로나 바이러스 발생 이후에 Netflix를 중심으로 OTT 상에서 K-콘텐츠 수요가 폭발적으로 늘어난 현재, 스튜디오의 평균 작업 속도가 OTT 플랫폼 콘텐츠 제작 수요를 따라잡지 못하고 있는 실정이어서, 외국의 사운드 스튜디오에서 사운드 후처리 작업도 다수 발생하는 상황이다. AI Sound Generation 기술이 사운드 스튜디오에 공급될 경우, 오디오 후처리 작업 속도에 파괴적 혁신을 일으킬 수 있을 것으로 기대된다. 특히, 전체 후반 작업 중, AI Sound Generation 기

**그림 5.** 메타버스에 대응되는 AI Sound Genration 개념도




술이 적용될 수 있는 Foley 생성과 효과음 생성은 전체 과정의 40~80% 정도의 시간이 소요되는 작업임을 감안하면, 이론상 생산성 향상은 60%~500%까지도 이루어질 수 있으나, 현실적으로는 30%~80% 정도일 것으로 기대하고 있다.

## 가상 공간의 재구현 : 메타버스 플랫폼을 위한 오디오 기술

현존하는 메타버스 플랫폼인 제페토, 로블록스, 게더타운 등에서는 비디오 객체와 어울리는 오디오 경험이 부재하거나 다소 부족한 상태에 있다. 메타버스 플랫폼의 특성상 1인 제작자에 의존하는 사용자 참여형 콘텐츠의 대중화가 전체 서비스의 성패를 좌우하는 중요한 것임에도 불구하고, 참여형 콘텐츠에서 다음과 같은 이유로 오디오의 활용이 제한된다.

- 오디오 제작에 대한 어려움 : 일반 사용자가 만드는 참여형 World나 Asset에서 배경 음악 이상의 효과음을 만들어 내는 작업 자체가 고도의 전문성을 요구하고, 객체 간의 상호작용 과정에서 발생하는 소리의 동기화 문제 등을 모두 해결하기 어렵다.
- 오디오에 대한 저작권 문제 : 불특정 다수의 1인 제작자에 의한 참여형 콘텐츠의 특성상 오디오의 활용성을 열어두기 때문에, 저작권에 위배되는 오디오가 사용될 경우 플랫폼이 그 저작권 문제를 해결해야 하는 법적 리스크를 지게 된다.
- 오디오에 대한 품질 문제 : 역시 참여형 콘텐츠의 특성상, 사용되는 오디오의 품질과 적합성을 관리할 수 없는 문제가 발생한다.

가우디오랩의 AI Sound Generation 기술은 이러한 메타버스 플랫폼에서, 플랫폼에서 용인하고자 하는 가이드라인에 따르면서도 Context에 대응되는 소리를 손쉽게 만들 수 있다. 개방적이나 관리 가능한 솔루션으로 활용될 수 있어, 플랫폼의 저작권 문제가 해결되고 품질상으로도 검증된 오디오를 사실상 무제한 생성하며 메타버스의 소리 경험이 혁신적으로 향상될 수 있을 것으로 기대된다.


**저자소개**

**전상배 가우디오랩 CSO**는 서울대학교 음향공학박사이자 약 20년 경력의 오디오 기술 연구자이다. 가우디오랩 합류 전 삼성전자 DMC연구소(현 삼성리서치)에서 입체 음향 관련 연구를 진행하는 책임 및 수석 연구원으로 재직하며 입체 음향 표준인 MPEG-H 3D Audio의 표준기술을 개발하며 삼성전자의 표준 IP 확보에 기여하였을 뿐만 아니라 스마트폰, TV, 사운드바, 홈시어터 등의 삼성전자 제품 차별화 IP를 다수 개발하였다. 현재는 가우디오랩 Chief Science Officer로 Spatial Audio 및 AI Audio 기술 연구개발을 책임지고 있다. 2021년 11월 IITP(정보통신기획평가원) 선정 ICT R&D 기술개발 우수연구자로 선정된 바 있다.