JDCS 디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 5, pp. 1151-1159, May. 2025

Check for updates

# 오디오 데이터셋의 다차원 분석과 시각화

오 희 원*
동국대학교 영상대학원 멀티미디어학과 Marte Lab. 연구원

# Multidimensional Analysis and Visualization of Audio Data

**Heewon Oh***
Researcher, MARTE Lab., Department of Multimedia, Dongguk University, Seoul 04620, Korea

**[요 약]**

본 연구는 대규모 오디오 데이터셋의 분석과 시각화를 위한 다차원 접근 방식을 제안한다. 기존의 메타데이터 기반 분석 방법은 확장성과 자동화 측면에서 한계가 있으며, 대용량 오디오 데이터의 효율적인 탐색을 저해한다. 이러한 문제를 해결하기 위해, 본 연구는 오디오 설명자 기반의 차원 축소 및 클러스터링 기법을 통합하고, 3차원 포인트 클라우드를 활용한 인터랙티브 시각화 기법을 도입하였다. 제안된 시스템은 t-SNE와 UMAP을 k-means 클러스터링과 결합하여 고차원 음향 데이터의 유사도 기반 구조를 직관적으로 파악할 수 있도록 지원한다. 또한 PCA 기반 색상 매핑과 샘플링 기법을 통해 시각적 해석력을 강화하였다. 해당 시스템은 기존 수동 주석 방식의 한계를 보완하며, 기계학습 기반 분석을 통해 효율적인 탐색 환경을 제공한다. 제안된 방식은 MIR, 사운드 디자인, 음향 데이터 분석 등 다양한 분야에 활용될 수 있다.

**[Abstract]**

This study proposes a multidimensional approach for analyzing and visualizing large-scale audio datasets. Traditional metadata-based analysis methods exhibit limitations in scalability and automation, which hinder efficient exploration of extensive audio datasets. To address these challenges, the proposed system integrates dimensionality reduction and clustering techniques based on audio descriptors and introduces an interactive 3D point-cloud visualization framework. By combining t-SNE and UMAP with k-means clustering, this approach facilitates intuitive understanding of structural patterns in high-dimensional acoustic data. PCA-based color mapping and representative data sampling are employed to enhance interpretability. The system overcomes the constraints of conventional tagging-based methods and enables efficient, interactive exploration through machine learning-driven analysis. This approach has potential applications in music information retrieval (MIR), sound design, and acoustic data visualization.

**색인어 :** 오디오 시각화, 클러스터링, 데이터 분석, 차원 축소, 기계 학습
**Keyword :** Audio Visualization, Clustering, Data Analysis, Dimensionality Reduction, Machine Learning

http://dx.doi.org/10.9728/dcs.2025.26.5.1151



**Received** 22 March 2025; **Revised** 09 April 2025
**Accepted** 23 April 2025

***Corresponding Author; Heewon Oh**
**Tel:** +82-70-8740-2738
**E-mail:** unohee.official@gmail.com

## Ⅰ. 서 론

인공지능 시대를 맞아 대량의 음향 데이터를 분류, 분석, 연구에 활용하는 사례가 점차 증가하고 있다. 그러나 기존 메타데이터(metadata) 기반 음원 분류 방식은 인간이 직접 레이블링을 수행해야 하므로 많은 시간과 비용이 소요되며, 레이블링되지 않은 음향 데이터를 즉각적으로 처리하지 못하는 구조적 한계를 지닌다. 더불어, 오디오 데이터셋(audio dataset) 생산 과정에 참여하는 주석자(annotator)들의 수동 주석(manual annotation) 방식은 오디오 데이터에 대한 주관적 편향과 확장성 부족 문제를 내포하고 있으며, 특히 다양한 음향 표현을 포함하는 대규모 오디오 데이터셋을 처리하는 데 있어 근본적인 취약성을 드러낸다[1]. 본 연구는 음향 데이터가 가진 음악적/비음악적 특성을 쉽게 파악할 뿐만 아니라, 다른 데이터 사이의 관계를 분석하기 위한 새로운 방법론으로 포인트 클라우드(point cloud) 기반 오디오 분석 프레임워크를 제안한다. 이는 t-SNE를 통한 국소적 구조 보존[2]과 UMAP 기반 전역적 최적화[3]와 같은 비지도 학습(unsupervised learning)기법을 계층적으로 적용하여 3차원 포인트 클라우드에 매핑한다. 이는 다차원 음향 특성의 잠재 공간(latent space)을 인간이 인지 가능한 수준의 저차원으로 압축하여 몰입감 있는 탐색 경험을 제공하는 데 목적이 있다. 뿐만 아니라 데이터 분석 / 시각화 파이프라인을 통하여 메타데이터가 부재한 비정형 오디오 데이터셋의 분석이 가능하며, 음향 이벤트 간 유사성을 직관적으로 파악할 수 있다. 이 방식은 기존의 메타데이터 기반 방식에 비해 처리 효율성과 분석의 직관성을 크게 향상시킨다. 특히, 차원 압축 기법은 저차원 데이터로 압축하는 데 목적이 있기 때문에 데이터의 군집 특성은 보존한 채 다른 형태의 시각화 방식과도 보완적인 구조를 만들 수 있다. 이를 통해 대규모 오디오 데이터셋에서의 저지연 인터랙티브 환경을 구현하고 실시간 탐색 효율성과 기계 학습(machine learning) 기반의 분석 정확도를 높일 뿐만 아니라, 이를 통해 새로운 음향 오디오 데이터셋을 생산하는 데 있어 노동 집약적인 기존 방식의 한계를 극복하는 중요한 기술적 진전을 제공할 것으로 기대된다. 이 연구는 음악 정보 검색(MIR), 사운드 디자인, 음향 분석 등 다양한 분야에서 효율적이고 실용적인 도구로 적용될 수 있는 가능성을 제시한다.

## Ⅱ. 선행 연구

### 2-1 기존 대형 오디오 데이터셋 접근법의 한계와 개선 방향

대규모 사운드 오디오 데이터셋을 활용하는 기존 접근 방식은 (1) 메타데이터 주석(annotation) 중심의 전통적 분석, (2) 기계 학습(machine learning)을 도입하되 오프라인(offline)·배치 프로세싱(batch processing)에 치중된 방식, (3) 시각화 단계에서 음향 특성을 충분히 반영하지 못한 사례 등으로 크게 구분할 수 있다. 메타데이터 기반 방법은 설정이 단순하고 직관적이라는 장점이 있으나, 대형 오디오 데이터셋 확장 시 레이블링 작업에 과도한 비용이 들어가고 주석의 주관적 편향 역시 문제로 지적되어 왔다[5]. 기계 학습 기법을 일부 도입하더라도 오프라인 학습이 주를 이루어 실시간 상호작용 측면이 약화되거나, 계산량 문제로 인해 상호작용형 시각화 구현이 제한적이다. 이러한 한계를 보완하기 위해 최근에는 오디오 설명자(audio descriptor)를 기반으로 실시간 분류·세분화·재합성을 시도하는 툴킷이 시도되고 있다. 특히 대규모 오디오 데이터셋에서 저지연 처리를 지원하고, 실시간 상호작용 인터페이스와 결합해 창작 및 연구에서 활용도를 높이려는 움직임이 활발히 이루어지고 있다. FluCoMa와 MuBu는 이러한 요구를 충족하기 위해 설계된 대표적인 도구들로, 각각 실시간 오디오 처리와 기계 학습을 결합한 접근 방식을 제공한다. 다음 절(2.2, 2.3)에서는 이를 대표하는 시스템을 간략히 살펴보고, 본 연구가 어떤 개선 방향을 제안하는지를 논의하고자 한다.

### 2-2 기존 연구

1) XLN Audio XO

XLN Audio XO는 스웨덴의 XLN Audio에서 개발한 오디오 플러그인으로, 방대한 샘플 라이브러리를 자동으로 정렬하고 표현하는 기능을 제공하는 드럼머신 / 시퀀서 소프트웨어이다. 기계 학습 기반 알고리즘으로 각 오디오 샘플의 스펙트럼(spectrum) 및 타임 도메인(time domain) 특징을 분석해 음색 유사도에 따라 오디오 샘플을 자동 분류하며, 이를 2차원 공간 형태로 시각화한다. 이를 통해 사용자 입장에서는 원하는 사운드를 빠르게 발견하고, 다양한 오디오 샘플을 한눈에 비교 및 평가할 수 있는 장점을 지닌다.


그림 1. XLN Audio XO의 사용자 인터페이스
**Fig. 1.** User interface of XO, XLN Audio

2) Ableton Live 12의 Sound Similarity Search

독일의 Ableton에서 개발된 Live 12는 DAW(Digital Audio

Workstation) 환경에 사운드 유사성 검색 기능을 내장하여, 프로젝트 내에서 음원을 재검색하거나 새로운 샘플을 탐색할 때 활용할 수 있도록 설계하였다. Sound Similarity Search 기능은 내부적으로 음색 분석 및 임베딩 기법을 결합해 음원들 간 유사도를 계산한 후, 관련도 높은 샘플을 사용자에게 추천해 주는 방식으로, 기존의 메타데이터 주석 기반의 브라우저 기능을 강화하는 방식으로 사용되었다.

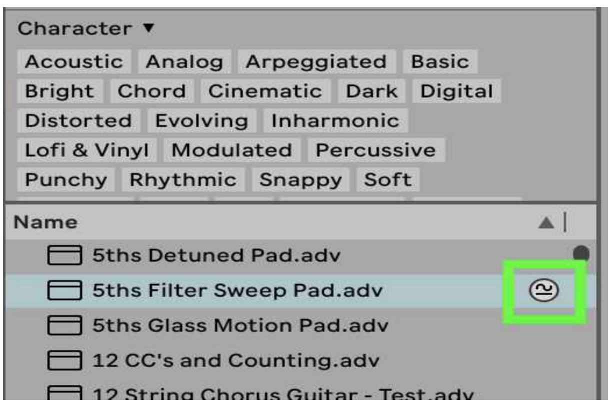


**그림 2.** 에이블톤 12의 오디오 유사성 검색
**Fig. 2.** Sound similarity search in Ableton 12

3) IRCAM MuBu(Multi-Buffer) 및 CataRT

IRCAM의 MuBu는 프랑스의 IRCAM(Institut de Recherche et Coordination Acoustique/Musique)에서 개발한 시스템으로, 오디오 신호뿐만 아니라 연속적인 분석값과 주석 정보를 함께 처리할 수 있는 구조를 제공함으로써 다층 구조 기반의 실험사운드 재합성 및 인터랙티브 조작을 지원한다. Max/MSP 환경에서 다층(multi-layer) 구조를 유지하는 멀티버퍼를 제공하여, 오디오 신호·설명자·마커·주석 등 복합 데이터를 통합적으로 다룰 수 있으며, 실시간 음원 재합성과 미세 조작을 지원한다. 이를 통해 음향적 특징이 유사한 소리를 그룹화하고 재합성함으로써, 다층적 사운드 텍스처를 직관적으로 설계할 수 있도록 하는 라이브러리이다.

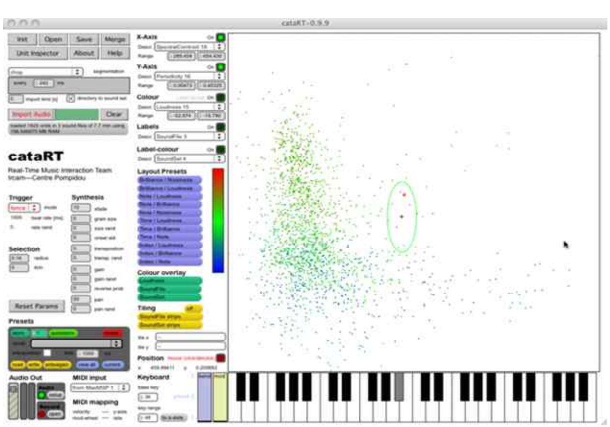

**그림 3.** IRCAM의 MuBu를 활용한 cataRT 인터페이스
**Fig. 3.** cataRT interface using IRCAM's MuBu

### 2-3 연구 방향 및 개선 사항

위와 같은 선행 연구들은 대형 오디오 데이터셋을 실시간으로 분석하고 시각화하는 데 있어 다양한 가능성을 제시하고 있다. XLN Audio의 XO는 방대한 샘플 라이브러리를 자동으로 분류 및 시각화하여 사용자 친화적 인터페이스를 제공하고 있으며, Ableton Live 12의 Sound Similarity Search는 DAW 환경에서 유사도 기반 탐색을 실시간으로 구현하여 창작 효율성을 높이고 있다. MuBu는 다층 구조 기반의 실험적 사운드 재합성과 인터랙티브 조작에 강점을 가진다. 그러나 XO의 경우, 드럼 샘플 분류 범위가 제한적이며, 몇 가지 색상으로 악기(스네어, 킥 드럼, 탐탐, 심벌 등)를 구분하는 데 그치는 등 기능이 국한된다. MuBu나 Flucoma와 같은 시스템은 Max/MSP 환경 내 기존 구조를 기반으로 작동하기 때문에, 분석 대상인 오디오 샘플(audio sample)뿐 아니라 오디오 설명자 데이터 또한 Max/MSP에서는 오디오 버퍼 또는 멀티버퍼 형태로 처리된다. 이러한 시스템은 주로 비동기적 처리보다는 실시간 처리를 우선하는 동기식 실행 환경에서 작동되며, 이로 인해 서로 다른 유형의 오디오 설명자를 동시적으로 분석하거나 다차원 데이터 구조를 저장·관리하는 데 한계가 있으며, 처리 과정에서 데이터 크기에 따라 병목 현상이 발생할 가능성도 존재한다. 또한 Max/MSP 환경 내에서 jsui와 같은 제한적인 인터페이스만을 지원함에 따라, 시각화 과정을 능동적으로 구성하거나 유연하게 확장하는 데 어려움이 있어 대형 오디오 데이터셋 분석 목적에는 부적합하다고 볼 수 있다. 이에 본 연구에서는 대규모 오디오 데이터셋을 효과적으로 시각화할 수 있는 방법론으로 비지도 학습 기반의 시각화 모델을 제안한다. 본 방법론은 오디오 데이터셋에서 추출된 고차원 특성을 3차원 포인트 클라우드 형태로 압축하여 표현함으로써, 기존의 2차원 평면 인터페이스뿐 아니라 AR(augmented reality) 및 VR(virtual reality) 같은 새로운 인터페이스나 관객 참여형 미디어아트 작업에서 더욱 직관적이고 효율적인 상호작용이 가능한 사운드 데이터 탐색 환경을 구축하는 것을 목표로 한다. 이와 같은 방식은 기존 선행 연구들이 제시한 장점들을 통합적으로 활용하면서도, 대규모 오디오 데이터셋의 분석 및 창작에 특화된 새로운 환경을 제공할 수 있을 것으로 기대된다. 또한 이러한 흐름은 딥러닝 기반의 공간 음향 분리 기법의 발전과도 방향성을 공유하며[4], 데이터 기반 오디오 해석 기술의 확장 가능성을 뒷받침한다.

## Ⅲ. 데이터 전처리 및 오디오 설명자 추출

### 3-1 사용 오디오 데이터셋

본 연구는 음원 분석 시스템의 성능을 평가하기 위해

Freesound 기반 공개 오디오 데이터셋인 FSD50K[6]를 사용하였다. FSD50K 오디오 데이터셋은 50,000개의 오디오 샘플과 전문가 검수를 거친 200개 이상의 음향 이벤트 클래스로 구성된 메타데이터를 포함한다. 오디오 데이터셋은 44.1kHz 샘플레이트(samplerate)로 제작되어 있으며, 음악적 및 비음악적 도메인(domain)을 아우르는 다양한 음향 스펙트럼을 포괄한다.

### 3-2 오디오 데이터셋 전처리 파이프라인

#### 1) 음원 표준화(Normalization)

본 연구에서는 스테레오 음원을 단일 채널로 다운믹스하여 분석에 활용하였다. 이는 음원의 공간적 방향성이나 위치 정보를 분석 대상에 포함하지 않으며, 주파수 기반 오디오 설명자(MFCC, 스펙트럴 중심, 크로마 등)의 추출을 통해 레이블 작업이 되지 않은 오디오 데이터셋의 음향적 유사성 및 관계를 파악하려는 본 연구의 목적에 부합한다. 특히 대부분의 스펙트럼 도메인 기반 오디오 설명자는 채널 간 평균 또는 합성된 파형으로부터 유사한 특성을 보인다. 이러한 오디오 설명자는 Flucoma와 같은 실시간 프로세싱 기반의 저지연 인터페이스 설계를 위해서 계산 효율을 높일 뿐만 아니라 데이터 일관성을 확보하기 위하여 모노 채널로 변환하였다. 뿐만 아니라 분석 단계에서는 불필요한 고음역대 정보를 줄이기 위하여 16kHz로 다운샘플링하였다. 마지막으로, 음원 전체를 −1 ~ +1 범위로 정규화함으로써, 음량 차이로 인해 발생할 수 있는 분석 편차를 최소화하였다. 이러한 표준화 과정은 데이터 품질을 균일하게 유지하여 모델 학습의 노이즈를 줄이고, 처리 속도를 높이는 데 기여한다.

#### 2) Onset 탐지(Onset Detection)

전처리된 오디오 데이터를 기반으로 소리의 시작 지점(onset)을 탐지하여 음원 구조를 보다 세밀하게 분석할 수 있다. onset 탐지는 LibROSA 라이브러리의 알고리즘을 사용하되, 표준화된 오디오 신호 기준 약 -40dBFS 이하의 에너지는 잡음으로 간주하여 제외하였다. 이 과정을 통해 오디오 신호 내에서 유의미한 이벤트 지점을 자동으로 식별하고, 이후 분할을 위한 기준점으로 활용한다.

#### 3) 오디오 분할(Segmentation)

onset 탐지에서 얻어진 지점을 바탕으로 오디오 데이터를 분할하되, 각 구간이 200ms 이상 되도록 설정하였다. 너무 짧은 구간(200ms 미만)의 분석 결과가 가진 음향 정보는 분석 결과의 정보량이 제한적이어서 유의미한 분석이 어렵다고 판단하여 제외한다. 분할된 각 오디오 유닛은 특징 추출 단계에서 독립된 분석 대상으로 처리되며, 필요 시 물리적인 파일로도 저장될 수 있다. 이렇게 생성된 오디오 조각들은 후속 모델 학습의 입력 데이터로 활용되며, 각 분할 구간에 대한 시작점(start)과 끝점(end)은 별도의 메타데이터(JSON 파일)에 기록되어 추후 분석 및 재구성에 이용된다.

### 3-3 오디오 설명자 추출(Audio Descriptor Extraction)

#### 1) 스펙트럴 중심(Spectral Centroid)

그림 4는 오디오 데이터가 가진 스펙트럴 중심 분포의 변화를 나타낸다. 세로축은 평균 스펙트럴 중심을 헤르츠(hz) 단위로 표현하며, 가로축은 오디오 샘플의 인덱스를 나타낸다. 그래프에서 보이는 변동은 다양한 오디오 샘플의 음색 차이를 시각적으로 보여주며, 특정 인덱스에서의 급격한 상승은 해당 샘플의 음색이 상대적으로 밝음을 의미한다. 이를 통해 오디오 데이터셋이 가진 음 높이의 분포도를 확인할 수 있다.

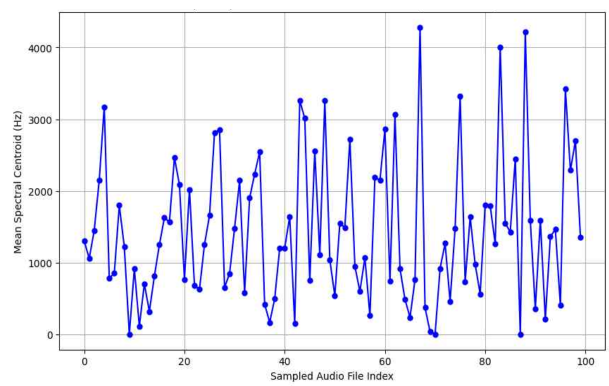


**그림 4.** 오디오 데이터셋 일부의 스펙트럴 중심 분포
**Fig. 4.** Sampled spectral centroid distribution of audio files

#### 2) 크로마 피쳐(Chroma Feature)

크로마 피쳐는 주파수 스펙트럼을 12개의 피치(pitch) 클래스로 나누어 분류하는 오디오 설명자이다. 이는 피치 관련 정보를 효과적으로 추출할 수 있어 음악적 콘텐츠의 하모니와 코드 분석에 유용하다. 반대로 크로마 피쳐 데이터가 명확하지 않은 경우 비음악적 데이터로 간주하는 것도 가능하다. 그림5는 오디오 데이터의 크로마 피쳐를 시각적으로 나타낸 것이다. 세로축은 음악의 12개의 피치 클래스를 나타내며, C부터 B까지의 음을 포함한다. 가로축은 시간 축을 나타내어 오디오 신호의 진행을 표현한다.

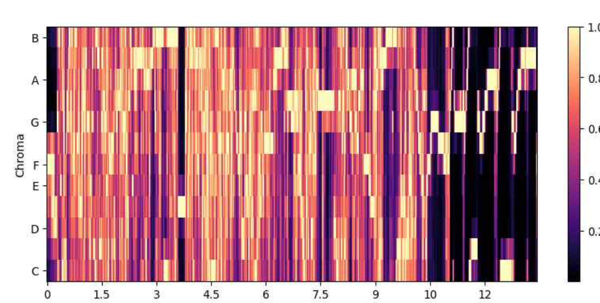


**그림 5.** 크로마 피쳐의 시각화
**Fig. 5.** Visualization of chroma feature

3) Mel-Frequency Cepstral Coefficients (MFCC)

MFCC는 음성 신호와 음악 신호 분석에서 가장 널리 사용되는 특징 추출 기법 중 하나이다. MFCC는 저주파수 대역에서 나타나는 음원의 형태적 특성과 에너지를 간단한 벡터로 압축하기 때문에, 다양한 오디오 분석 및 분류 작업에서 핵심 오디오 설명자로 널리 활용된다.

### 3-4 오디오 설명자의 활용

본 연구는 음악적/비음악적 콘텐츠의 주파수 기반 특성을 파악하기 위해 MFCC, 스펙트럴 중심, 크로마 피쳐, 세 가지 오디오 설명자를 조합하여 활용하였다. MFCC는 음색과 관련된 주파수 대역 분포를 효율적으로 요약하며, 스펙트럴 중심은 주파수의 에너지 중심 위치를 나타내어 밝기나 명료도와 같은 감각적 특성과 연관된다. 크로마 피쳐는 전체 스펙트럼을 12개의 피치 클래스로 투영하여 코드, 하모니 등 음악적 구성요소를 시각화하는 데 유용하다. 특히 크로마 피쳐의 경우 피치 특성이 뚜렷하지 않을 경우 비음악적 데이터로도 분류 가능하다. 이처럼 세 오디오 설명자는 서로 보완적인 정보를 제공하며, 다차원 오디오 설명자의 대표 집합으로 활용되었다. 또한 이러한 오디오 설명자의 집합은 차후 진행할 '데이터 시각화' 단계에서 고차원 임베딩과 포인트 클라우드 시각화 기법을 적용하는 출발점이 된다. 데이터 시각화 과정에서는 이들 오디오 설명자와 비지도 학습 기법을 결합하여, 대규모 오디오 데이터셋 내 음향 유사도와 구조적 패턴을 직관적으로 드러내는 시각화 방법론을 구체적으로 논의하고자 한다.

## Ⅳ. 데이터 시각화

### 4-1 데이터 차원 압축 및 샘플링 분석의 필요성

다수의 오디오 유닛 간 관계를 이해하기 위해 메타데이터 기반 수동 주석 방식을 사용할 경우, 주석자들 사이의 인지적 편차로 인해 동일 음향 이벤트에 대한 주석의 일관성이 58%로 제한되며, 이는 데이터 분석의 직관성과 객관성을 저해하는 것을 확인할 수 있다[5]. 분석된 오디오 설명자 데이터를 결합하여 분석하지 않는다면, 다양한 유형의 사운드 데이터는 하나의 설명자로는 효과적으로 표현되기 어려우며, 이를 보완하기 위한 설명자 간의 상호보완적 조합이 필요함을 실험을 통해 확인하였다. 따라서 본 연구에서는 MFCC, 스펙트럴 중심, 크로마 피처의 조합을 통해 이러한 상호보완성을 실현하고자 하였다. 그러나, 모든 오디오 설명자 데이터를 결합한 고차원 데이터를 분석에 사용할 경우 직관적인 시각화가 어렵고 기존 저차원 분석 기법의 성능이 급격히 저하되어 분석 효율이 급격히 떨어지는 문제가 있다. 따라서 본 연구에서는 고차원 데이터의 차원을 축소하는 기법과 포인트 클라우드 시각화 기법을 결합하여 고차원 오디오 설명자 데이터의 복잡성을 줄이고, 사운드 유닛 간 상호 유사도와 군집 양상을 직관적으로 드러내는 방법을 제안한다. 이를 통해 대규모 오디오 데이터셋(audio dataset)의 구조적 특징을 더욱 명확하게 해석할 수 있을 것으로 기대한다.

### 4-2 t-SNE를 이용한 차원 압축

t-SNE는 복잡한 고차원 데이터를 저차원 공간으로 투영하여 시각화하는 대표적 기법이며 데이터의 국소 구조를 보존하면서도 각 데이터 포인트 간의 상대적 거리를 직관적으로 파악할 수 있도록 돕는다. 본 연구에서는 세 가지 오디오 설명자('MFCC 평균', '스펙트럴 중심 평균', '크로마 평균')를 하나의 특징 벡터로 결합한 뒤, 불완전하거나 형식이 불일치한 벡터를 제외하고 나머지를 임의의 3차원 벡터로 압축하여 임베딩(embedding)하였다. 그 결과, 각 오디오 유닛은 임베딩 값을 토대로 3차원 공간에서 하나의 점으로 표현되며, 이 점들의 상대적 위치를 통해 음향 이벤트들의 군집 양상을 시각적으로 파악할 수 있게 된다.

### 4-3 3차원 임베딩의 시각화 및 컬러 매핑

최종적으로 얻어진 3차원 t-SNE 임베딩은 단순히 좌표 값으로만 시각화되는 것이 아니라, 추가적인 음향 특성을 반영한 색상 정보를 통해 보다 풍부하게 표현된다. 본 연구에서는 원하는 파라미터에 따라 다음과 같은 색상 매핑 기법을 적용하였다.

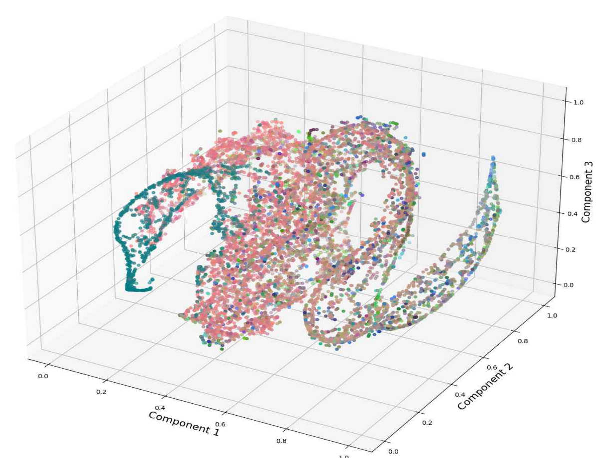


**그림 6.** PCA를 적용한 크로마 피쳐 기반의 색상으로 표현한 3차원 t-SNE 시각화

**Fig. 6.** 3D t-SNE visualization colored by chroma features reduced via PCA

먼저, 12음계를 표현하는 크로마 특징과 같은 고차원 정보를 가진 오디오 설명자의 경우, PCA 기법(Principal Component

Analysis)을 통해 데이터를 3차원으로 축소한 후, 각 주성분을 R, G, B 채널에 대응시켜 RGB 색상으로 매핑하였다(그림 6). 이러한 방식은 각 음향 이벤트의 주요 피치 클래스를 시각적으로 강조하여, 색상을 통해 음향의 구조적 특성을 직관적으로 파악할 수 있도록 돕는다. 이를 통해 음악적 구성 요소(예: 코드, 하모니 등)를 시각적으로 표현하고, 사용자가 복잡한 음향 특성을 쉽게 이해하고 탐색할 수 있게 만들어준다. RGB 색상은 특히 음향의 높은 피치나 저음 피치를 직관적으로 구별할 수 있게 해주며, 음악적 패턴을 더 쉽게 파악하도록 돕는다. 반면, 스펙트럴 중심과 같은 단일 스칼라 값을 가지는 파라미터의 경우는 viridis 컬러맵을 적용하여 시각화하였다(그림 7). 따라서, 음의 높낮이가 오디오 스펙트럼으로 수렴하여 미세한 색상 변화를 통해 음색의 밝기나 특성을 직관적으로 전달한다.

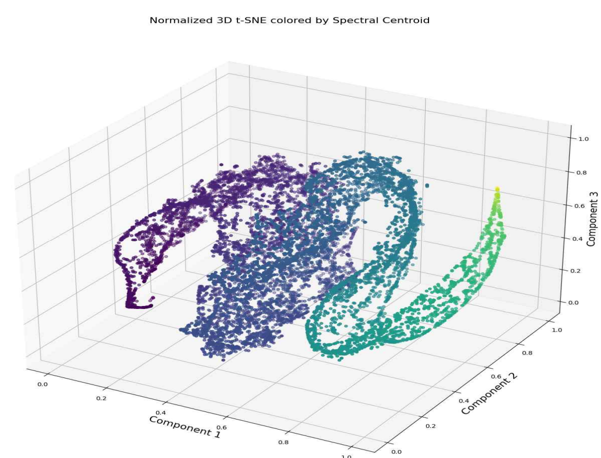


**그림 7.** 스펙트럴 중심 값과 viridis 컬러맵을 적용한 3차원 t-SNE 시각화

**Fig. 7.** 3D t-SNE visualization colored by spectral centroid values using the viridis color map

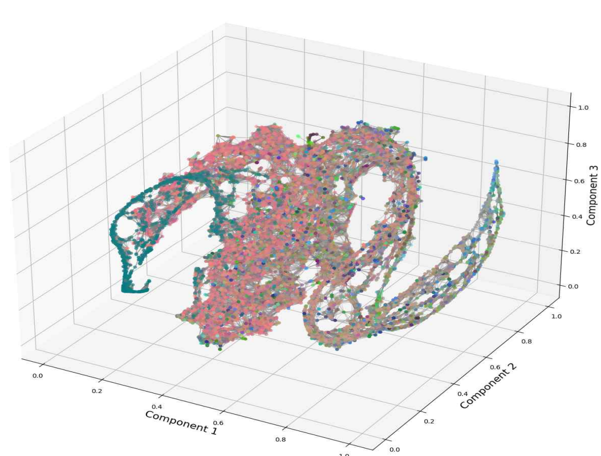


**그림 8.** 스펙트럴 중심 유사도에 따른 t-SNE 시각화 및 선분 연결

**Fig. 8.** t-SNE visualization with edges based on spectral centroid similarity

또한, 복수의 설명자 사이의 관계를 나타내기 위해 t-SNE 임베딩 공간 상에서 각 오디오 유닛 간의 유클리드 거리가 일정 임계치(예: 0.1) 이하이며, 동시에 스펙트럴 중심 정규화 값의 차이가 일정 수치(예: 0.2) 이하인 경우, 이들 간의 유사성을 강조하기 위하여 회색 선분(엣지)을 연결하였다. 이러한 선분은 음향 특성이 유사한 오디오 유닛 간의 관계를 직관적으로 부각시키는 역할을 한다(그림 8).

### 1) 대형 오디오 데이터셋 t-SNE 시각화의 한계점

앞선 테스트에서는 10,000개의 오디오 샘플에 대해 t-SNE 시각화를 수행하였으나, 실제로 FSD50K와 같이 약 50,000개의 파일과 10GB 이상의 용량을 가진 대규모 오디오 데이터셋을 시각화할 경우, 데이터 양의 증가로 인해 t-SNE의 계산 효율성이 급격히 저하되는 문제가 발생하였다. 특히 고차원 공간에서는 모든 데이터 포인트 간 유클리드 거리가 상대적으로 수렴하면서, 유사도 기반의 분별력이 저하되는 현상이 관찰되었다. 그 결과, 시각화된 오디오 유닛 간의 분포가 과도하게 밀집되어 보이거나 명확한 군집 경계가 드러나지 않는 문제가 발생하였다(그림 9). 또한, 모든 포인트 간 유사도 기반 시각화를 수행할 경우, 이론적으로는 $O(n^2)$ 수준의 연산이 요구되며, 이는 대규모 오디오 데이터셋에서 실시간 시각화에 부담을 줄 수 있다. 본 연구에서는 이를 조건 기반 필터링 방식으로 제어하였으며, 추후에는 KD-Tree 기반 근접 탐색 기법을 적용함으로써 연산 효율을 더욱 향상시킬 수 있을 것으로 기대된다.

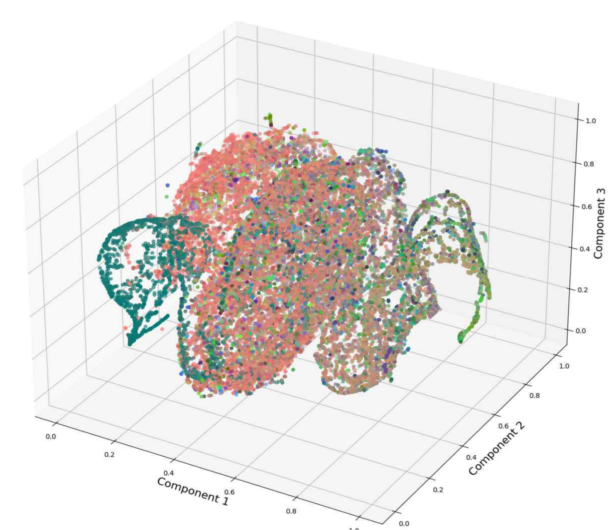


**그림 9.** FSD50K 오디오 데이터셋의 3차원 t-SNE 시각화

**Fig. 9.** 3D t-SNE visualization of FSD50K dataset

### 2) UMAP/k-means Clustering을 이용한 데이터 샘플링

본 연구는 FSD50K의 t-SNE를 활용한 임베딩 결과가 매우 높은 밀집도를 형성한다는 점에 주목하여, 전체 오디오 데이터셋 중 무작위로 선택한 2,500개의 오디오 유닛을 대상으

로 UMAP 기반 3차원 임베딩을 수행하였다. 이후 해당 임베딩 결과에 대해 k-means 클러스터링을 적용함으로써 국소적인 군집 구조를 분석하고 시각화하였다. 이러한 방식은 전체 구조를 대표값으로 축약하기보다는, 고차원 공간의 밀집된 일부분을 탐색적으로 접근함으로써 오디오 데이터셋의 잠재적 패턴을 효과적으로 드러낸다(그림 10).

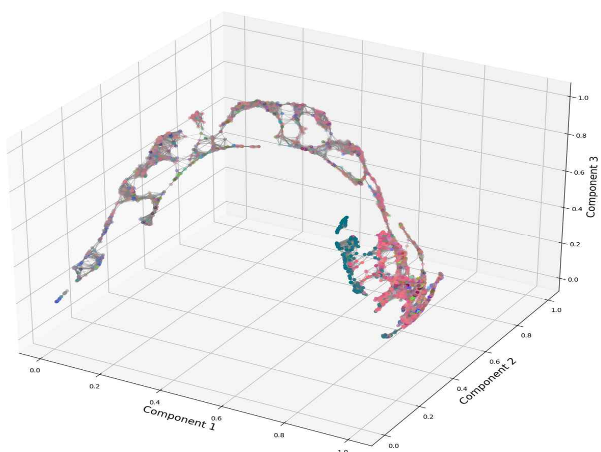


**그림 10.** 스펙트럴 중심 유사도를 이용한 UMAP 기반 샘플링 오디오 데이터셋의 3차원 시각화
**Fig. 10.** 3D visualization of UMAP-sampled dataset with spectral centroid similarity edges

무작위 샘플링 기반 시각화 전략은 특히 전체 데이터 공간의 전반적 분포를 균형 있게 탐색할 수 있다는 점에서 유의미하며, 연산 자원 소모를 줄이면서도 고차원 오디오 설명자 간의 의미 있는 구조와 군집 특성을 직관적으로 관찰할 수 있는 방법이다. 또한 본 시스템은 향후 확장을 고려하여, t-SNE 기반의 2차원 임베딩 공간을 탐색 지도(map)처럼 활용하고, 사용자 커서가 위치한 국소 영역을 중심으로 UMAP 및 k-means를 실시간으로 수행하는 저지연(low-latency) 인터페이스로 확장하는 방향을 고려하고 있다. 이를 통해 사용자는 전체 구조를 전역적으로 파악함과 동시에, 관심 영역에 대한 세부 구조를 3차원 시각화 및 오디오 재생을 통해 분석할 수 있다. 이러한 실시간 상호작용 기반 국소 시각화 인터페이스는 고차원 오디오 데이터셋의 탐색과 창작 워크플로우를 유기적으로 연결하는 기반이 될 것이다.

## V. 차원 압축 기법의 유효성 검증

본 연구에서는 t-SNE 기반 클러스터링과 UMAP 및 k-means를 결합한 국소 샘플링 구조 간의 차이를 정량적으로 평가하기 위해, 실루엣 스코어(Silhouette score), 유클리드 거리(Euclidean distance), 그리고 임베딩 수행 시간(Embedding time)의 세 가지 지표를 사용하였다. 이때 측정된 시간은 시각화 과정을 제외하고, 차원 축소 및 클러스터링에 소요된 시간만을 기준으로 하였다(표 1).

**표 1.** 차원 축소 기법의 데이터 유효성 검증 지표
**Table 1.** Validation metrics for dimensionality reduction methods

| Method | Silhouette score | Euclidean distance (Mean, Std) | Embedding time (s) |
|---|---|---|---|
| t-SNE | 0.293 | 0.53 (±0.21) | 105.20 |
| UMAP | 0.375 | 0.62 (±0.28) | 28.7 |

### 5-1 평가 지표

#### 1) 실루엣 스코어

클러스터 내 응집도(cohesion)와 클러스터 간 분리도(separation)를 동시에 반영하는 실루엣 스코어를 산출하여 두 기법이 데이터를 얼마나 효과적으로 구성하는지를 정량적으로 측정하였다. 실루엣 스코어가 높을수록 각 클러스터 내부의 데이터 포인트들이 밀집되어 있고, 클러스터 간 경계가 명확함을 의미한다. 평가 결과 t-SNE 방식은 0.293의 실루엣 스코어를 보인 반면, UMAP 방식을 사용하여 일부를 샘플링하는 경우는 0.375로 나타나, 전반적인 군집화 품질에서 UMAP이 우수함을 확인할 수 있었다.

#### 2) 유클리드 거리 분포

고차원 데이터와 저차원 임베딩 사이의 유클리드 거리 분포(평균 및 표준 편차)를 분석하여, 각 기법이 데이터의 국소 및 전역 구조를 얼마나 효과적으로 보존하는지 평가하였다. t-SNE의 경우, 평균 유클리드 거리가 0.53(±0.21)로 측정되었으나, UMAP 기반 방법은 0.62(±0.28)로 다소 높은 값을 보임으로써 전역적 거리 분포를 보다 효과적으로 유지하고 있음을 시사한다.

#### 3) 임베딩 처리 시간

본 연구에서는 각 기법의 처리 효율성을 비교하기 위해, t-SNE 및 UMAP 기반 임베딩 과정에서 소요된 실제 연산 시간을 측정하였다. 본 비교는 시각화 렌더링 및 엣지 연결과 같은 후처리 단계를 제외하고, 차원 축소 및 클러스터링(UMAP + k-means) 과정에 한정하여 측정되었다. 그 결과, FSD50k 오디오 데이터셋을 대상으로 한 실험에서 t-SNE의 평균 임베딩 수행 시간은 약 105.2초, UMAP 기반 샘플링 기법은 약 28.7초로 측정되었으며, UMAP 방식이 약 72.7%의 시간 절감을 보였다. 이는 대규모 오디오 데이터셋을 실시간으로 분석하거나 상호작용 기반 시스템에 적용하는 데 있어 중요한 성능 개선 요소로 작용할 수 있다.

### 5-2 교차 유효성 검증

본 연구에서는 FSD50K 오디오 데이터셋의 오디오 메타데이터를 활용하여 각 데이터 포인트의 주석 정보를 기반으로 UMAP 임베딩 결과를 2차원 공간에 투영하고, 주요 주석 그룹별 중심 좌표를 계산하여 클러스터링 구조의 일관성을 분석하였다. 예를 들어, 특정 환경음(예: "rain", "wind")이 특정 영역에 집중적으로 위치하는지를 평가함으로써, 모델이 학습한 잠재 공간이 실질적인 사운드스케이프 특성과 얼마나 일치하는지를 분석한다(그림11). 본 검증 과정은 비지도 학습 기반 임베딩 및 클러스터링 결과가 메타데이터 기반 분류와 부합하는지를 평가하는 중요한 절차로 작용한다. 결과적으로, 본 연구에서 제안한 국소 샘플링 기반 클러스터링 방법은 내부 동질성과 도메인 관련성을 유지하면서도 고차원 오디오 데이터를 효과적으로 구조화할 수 있음을 확인하였으며, 오디오 데이터셋의 탐색 및 활용 가능성을 크게 향상시킬 수 있음을 기대할 수 있다.

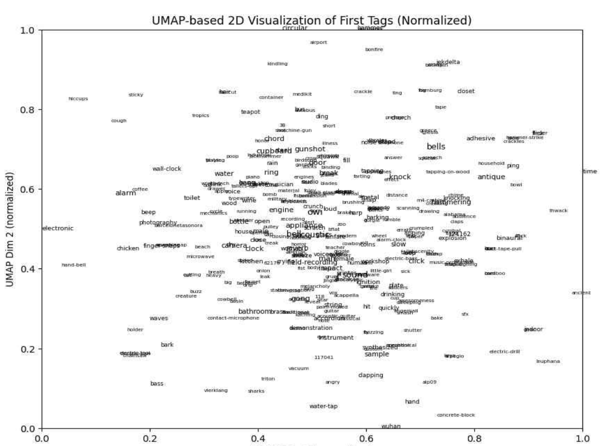


**그림 11.** UMAP으로 임베딩한 오디오 dataset의 메타데이터 기반 2차원 시각화

**Fig. 11.** 2D UMAP visualization of the audio dataset clustered by metadata labels

## Ⅵ. 결 론

본 연구는 대규모 오디오 데이터셋의 복합적 구조를 효과적으로 해석하고 탐색하기 위하여 차원 축소와 클러스터링 기법을 결합한 시각화 방법론을 제안하였다. 기존 메타데이터 기반 분석 시스템이 가진 주관적 편향과 확장성 부족의 문제를 극복하기 위해, 오디오 설명자 기반의 비지도 학습 접근 방식을 도입하였으며, 이를 통해 고차원 데이터의 복잡한 관계를 효과적으로 분석하고 시각화할 수 있는 시스템을 구현하였다. 연구의 주요 성과는 세 가지로 요약된다. 첫째, t-SNE와 UMAP을 활용한 차원 축소 기법은 고차원 오디오 데이터의 구조적 특성을 효과적으로 보존하면서, 데이터 포인트 간의 상호 유사도와 군집 양상을 직관적으로 표현할 수 있음을 확인하였다. 특히, UMAP 기반의 국소 샘플링 방식은 처리 시간 단축과 유클리드 거리 분포의 보존 측면에서 우수한 성능을 보여 대규모 오디오 데이터셋에도 적합함을 입증하였다. 둘째, k-means 클러스터링과 결합한 샘플링 기법은 데이터의 양을 적절히 줄이면서도, 주요 군집 특성을 유지하여 실시간 탐색이 가능한 시각화 시스템을 구현하는 데 기여하였다. 셋째, 오디오 설명자 간 추가적인 관계를 효과적으로 드러내기 위해, 12차원 크로마 피쳐에 대해 주성분분석(PCA)을 적용하여 3차원 RGB 색상으로 매핑하고, 스펙트럴 중심의 유사성을 기반으로 유클리드 거리 임계치 내에서 선분(엣지)으로 연결하는 기법을 도입함으로써, 음향 특성의 미세한 차이를 시각적으로 부각시켰다. 이러한 시각화 기법은 음향 분석, 사운드 디자인 등 다양한 분야에서 활용될 수 있는 실질적인 도구를 제공하며, 대규모 오디오 데이터셋의 복합적 구조와 내재된 관계성을 명확히 해석할 수 있는 새로운 가능성을 제시한다. 그러나 동시에, 본 연구는 다음과 같은 한계점을 내포한다. 첫째, 차원 축소 및 클러스터링 과정에서 활용된 오디오 설명자(MFCC, Spectral Centroid, Chroma Feature)는 음향의 일부 측면만을 반영하며, 감정, 음질, 공간 정보와 같은 고차원적 음향 특성은 충분히 포착하지 못한다. 둘째, FSD50K 및 FSDnoisy18k 데이터셋에 기반한 분석 결과는 특정한 레이블링 구조와 잡음 환경에 종속될 가능성이 있어, 다른 형태의 데이터셋에 일반화할 때 성능 저하가 발생할 수 있다. 셋째, 본 연구에서 제안한 시각화 시스템은 인터랙티브 탐색에 적합하도록 설계되었으나, 대규모 실시간 환경에서의 연산 성능과 확장성에 대한 정량적 검증은 충분히 수행되지 않았다. 이러한 한계는 향후 더 다양한 오디오 특성과 사용자 환경을 고려한 시스템 개선 및 고도화 연구를 통해 보완될 필요가 있다. 향후 연구에서는 다양한 오디오 설명자 데이터를 일반적인 오디오 버퍼 대신 텐서(tensor)로 변환하므로써 PyTorch와 같은 프레임워크를 사용하여 GPU 기반 가속 컴퓨팅 환경에서 실시간 분석 및 처리를 지원할 수 있을 것으로 기대된다. 이러한 확장은 기존 선행 연구들이 제시한 장점들을 통합하고, 대규모 오디오 데이터셋에 특화된 효율적이며 창의적인 분석 및 창작 환경을 구현하는 데 기여할 수 있을 것이다.

## 참고문헌

[1] J. Schlüter and T. Grill, "Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Spain, pp. 121-126, 2015.

[2] L. Maaten and G. Hinton, “Visualizing Data Using t-SNE,” *Journal of Machine Learning Research*, Vol. 9, pp. 2579-2605, 2008.

[3] L. McInnes, J. Healy, S. Nathaniel, and G. Lukas, “UMAP: Uniform Manifold Approximation and Projection,” *Journal of Open Source Software*, Vol. 3, No. 29, 2018. http://doi.org/10.21105/joss.00861

[4] F. Lluís, N. Meyer-Kahlen, V. Chatziioannou, and A. Hofmann, “Direction Specific Ambisonics Source Separation with End-to-End Deep Learning,” *Acta Acustica*, Vol. 7, 2023.

[5] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, “A Tutorial on Onset Detection in Music Signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 13, No. 5, pp. 1035-1047, 2005.

[6] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, “FSD50K: An Open Dataset of Human-Labeled Sound Events,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 829-852, 2021. http://doi.org/10.1109/TASLP.2021.3133208



**오희원(Heewon Oh)**

2022년 : 동국대학교 영상대학원 (컴퓨터음악석사)

2013년～2016년: Goldsmiths, University of London, BSc Music Computing 전공

2020년～2022년: 동국대학교 영상대학원 멀티미디어학과 컴퓨터음악전공 석사

※관심분야 : 머신러닝, 소리시각화, 컴퓨터음악, 실시간 사운드 프로세싱