The Transactions of the Korea Information Processing Society
TKIPS Letter Paper Vol.14, No.1, pp.48~52

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2025.14.1.48

Multi-Time Segment Peak-based Audio Fingerprinting

Son U Hyun† · Junyoung Heo††

다중 시간 세그먼트 피크 기반 오디오 지문 생성

손 유 현† · 허 준 영††

ABSTRACT

This paper explores methods to enhance the accuracy of audio fingerprinting technology. Since the early 2010s, audio fingerprinting has been actively researched for applications such as music and media retrieval. More recently, advancements in deep learning have spurred efforts to improve the accuracy of this technology. In this study, we propose an approach to enhance audio fingerprinting accuracy by introducing specific steps and techniques into the process. The first step involves detecting key peaks and generating audio fingerprints. This approach simplifies computational processes and achieves robust signal recognition even in noisy environments by leveraging the intensity of prominent frequency peaks. The second step involves generating audio fingerprints based on multi-time segment peaks. By extracting peaks from different audio segments, this method creates fingerprints that comprehensively reflect the characteristics of various temporal sections. This approach captures the temporal features of the audio in greater detail by incorporating segments of varying lengths. Comparative analysis reveals that while the proposed method maintains a similar processing time to conventional approaches, it delivers significantly higher accuracy.

Keywords : Audio Fingerprint, Multiple Time Segments

요 약

본 논문은 오디오 지문 기술의 정확도를 향상시키기 위한 방법을 다룬다. 2010년대 초반에는 오디오 지문 기술을 활용한 음악 및 미디어 검색 연구가 활발히 이루어졌으며, 최근에는 딥러닝을 활용하여 기술의 정확도를 높이는 연구가 진행되고 있다. 본 논문에서는 오디오 지문 기술에 특정 단계와 기법을 추가함으로써 정확도를 개선할 수 있는 방안을 제안한다. 먼저, 주요 피크를 탐지하여 오디오 지문을 생성하는 방법을 제시한다. 이 방법은 계산 과정이 간단하며, 특정 주파수 피크의 강도를 반영함으로써 잡음이 많은 환경에서도 신호를 정확히 인식할 수 있다. 다음으로, 다중 시간 세그먼트 피크 기반 오디오 지문 생성 방법을 소개한다. 이 기법은 오디오의 다양한 구간에서 피크를 추출하여, 여러 구간의 특성을 모두 반영한 지문을 생성한다. 이를 통해 다양한 길이의 세그먼트를 활용하여 오디오의 시간적 특성을 더욱 상세히 반영할 수 있다. 제안된 방법으로 추출한 오디오 지문을 기존 방식과 비교한 결과, 처리 시간은 유사했으나 정확도가 더 높은 것을 확인할 수 있었다.

키워드 : 오디오 지문, 다중 시간 세그먼트

1. 서 론

2000년대 이후, 전자 기기와 그에 따른 음악, 미디어의 보급률이 급증하며 음악과 미디어의 저작권 문제는 꾸준히 제기되어 왔다. 그리고 최근 몇 년은 스마트폰과 태블릿과 같은 모바일 기기의 보급률이 급증하며, 미디어 콘텐츠의 소비 방식에도 변화를 불러일으켰다. 그 변화의 중심에 있는것이 바로 OTT(Over-The-Top) 서비스이다. OTT 서비스가 변화의 중심이 되며 콘텐츠 제작자들은 새로운 기회가 생기고, 이에 따라 퀄리티 높은 미디어 콘텐츠가 다양하게 제작되고 있다. 이에 따라, 콘텐츠의 품질이나 시청률을 측정하고, 저작권을 보호할 기술의 필요성이 대두되었다.

기존의 방식은 TV방송에 초점이 맞추어져 있기에 OTT 서비스의 미디어 콘텐츠를 측정하고 보호하는 데에는 한계가 있었고, 그에 따라 다른 방식이 필요했는데, 그에 따라 연구되는 기술이 특정 오디오 파일을 이용하는 방식이다.

특정 오디오 파일을 식별하고 검색하는 기술은 저작권 보호, 방송 모니터링, 음악 검색, 광고 분석 등 다양한 산업에서 핵심적인 역할을 하고 있다. 본 논문에서 오디오 식별 및 검색 방법으로 소개할 기술은 오디오 지문 기술이다. 오디오 지문 기술은 음악과 미디어의 발전에 따라 2010년대 초반에 활발히 연구되어왔다. 오디오 지문 기술을 이용하여 음악을 요약하는 기술 [1]이나, 음악을 쿼리로 사용하여 singing이나 humming으로만 음악 검색을 할 수 있는 기술 [2] 등 다양한

※ 본 연구는 한성대학교 교내학술연구비 지원과제임.
† 준 회 원 : 한성대학교 컴퓨터공학과 석사과정, uhyun1517@gmail.com, https://orcid.org/0009-0007-5468-0987
†† 정 회 원 : 한성대학교 컴퓨터공학부 교수, jyheo@hansung.ac.kr, https://orcid.org/0000-0001-6407-6678
Manuscript Received : December 18, 2024
Accepted : December 27, 2024
* Corresponding Author : Junyoung Heo(jyheo@hansung.ac.kr)

방면으로 연구되고 사용되고 있다. 오디오 지문 연구에서 꾸준히 문제점으로 짚이고 있는 점은 잡음이 많은 환경에서의 정확성과 강인함인데, 최근에는 딥러닝을 도입하여 오디오 지문 기술의 정확성이 더욱 올라가고 있는데, Convolutional Neural Networks(CNN)을 활용한 오디오 인식 모델은 기존의 특징 추출보다 높은 성능을 보이고 있다. 하지만 본 논문은 딥러닝을 도입한 방법이 아닌, 기존의 오디오 지문 기술에 새로운 조건과 아이디어를 도입하여 정확도를 올릴 실용적인 방법을 제안한다.

2. 관련 연구

오디오 지문 기술은 일부 특징을 추출하여 저작권 데이터베이스를 구축해 두고 이와 비교함으로써 저작권 위반 여부를 확인하는 기술이다. 대표적인 연구로는 Shazam [3]이 있다. Shazam은 음악 식별 기술로, 음악별로 고유한 오디오 지문을 생성하여 음악 식별에 도움을 주고, 그 정보들을 토대로 어떤 음악이 성행하는지 파악하고, 사용자가 음악을 검색하는 것에 도움을 준다.

Wang [4]은 주파수와 시간 데이터에서 특징적인 피크를 추출하여, 이를 고유한 오디오 지문으로 사용한다. 이 지문은 데이터베이스에서 해시 테이블을 이용해 빠르게 검색이 가능하다.

오디오 지문 기술의 대표적인 문제점에는 잡음에 취약하다는 점이 있는데, 2010년대에 그에 관련된 연구들이 활발히 이루어졌다. TV 광고 식별을 위한 연구가 있는데 오디오 지문 기술을 이용하며, 다양한 잡음에도 정확성을 유지하기 위해 Constant Q [5] 방법을 사용하여 오디오 지문의 강인함을 높여준다.

오디오 지문 기술을 시청각 장애인용 방송을 위해 활용하는 연구 [6]도 있다. 국민의 90%가 사용하는 스마트 기기를 주 HW기기로 활용하여 방영되는 방송의 오디오 지문 정보를 통해 방송을 인식하고, 그에 따른 시청각 장애인들에게 필요한 자막 정보 및 화면해설 정보를 동기화 시킴으로써 스마트 기기의 스크린이나 이어폰으로 소비할 수 있게 해준다.

앞서 소개한 Shazam과 같은 오디오 지문 알고리즘도 성능은 우수하지만, 데이터 크기와 속도에 단점이 있어, 그것을 임베딩 기법을 통해 지문 데이터 크기를 줄이고 속도를 향상시키는 연구 [9]도 있다. LLE 기법을 통해 오디오 지문 크기를 감소시키고, DTW를 적용함으로써 검색 정확도도 기존과 유사한 성능을 유지하였다.

오디오 지문 기술에 신경망을 도입한 연구 [10]가 있는데, 오디오 파일들을 효과적으로 분류하기 위해 오디오 지문을 활용하고, 그것에 LSTM 알고리즘을 더해 이진화된 MFCC 데이터를 입력으로 사용해 오디오를 분류하여 높은 분류 정확도를 나타내었다.

서론에서 언급했던 CNN 기법을 이용한 음성, 음악 분류 연구 [11]도 있는데, 5초 단위의 오디오 데이터셋을 CNN과 전이학습을 이용하여 높은 정확도로 음성과 음악을 분류하였다.

이 기술들 외에도, 최근 딥러닝을 이용하여 오디오 지문의 정확성을 높이는 연구가 이루어지고 있다. CNN을 활용한 모델들은 기존의 연구보다 높은 성과를 보이고 있으며, 이를 통해 다양한 분야에서 오디오 지문 기술이 활용되고 있음을 알 수 있었다.

3. 오디오 지문 생성 기법

3.1 오디오 지문 기술

오디오 지문 기술은 오디오 파일의 식별, 검색 등에 사용되는 기술로, 음악 방송, 저작권 보호, 광고 모니터링 등 다양한 응용 분야에서 중요한 역할을 하는 기능이다. 오디오 지문 기술은 말 그대로 일종의 지문을 생성하는 기술인데, 오디오 파일의 고유한 특성을 추출하는 방식으로 생성한다.

오디오 지문 기술은 일반적으로 두 가지 단계로 나뉘는데, 첫 번째는 특징 벡터를 추출하는 단계이고, 두 번째는 추출한 특징 벡터를 데이터베이스에서 검색, 비교하는 단계이다. 이 과정에서 중요한 점은 다양한 신호 처리 기법과 알고리즘을 이용하여 왜곡에 강인한 특징이나, 잡음을 추출하는 것이다.

3.2 피크 기반 오디오 지문 생성

피크 기반 오디오 지문 생성이란 도메인에서 중요한 피크를 탐지하여 오디오 지문을 생성하는 방법이다. 이 방법은 스펙트럼 피크를 이용하여 오디오의 특징을 추출하는데, 생성 과정은 다음과 같다.

첫째, 오디오 신호를 청크로 분할한다. 이때 청크 길이는 주어진 세그먼트 길이를 샘플링 레이트로 곱한 값이 된다. 본 연구에서는 샘플링 레이트를 22050Hz의 오디오 데이터를 사용한다. 예를 들어 세그먼트 길이가 10초, 샘플링 레이트가 22050Hz일 경우, 청크 길이는 220500가 되는 것이다. 청크 크기를 키우면 주파수 해상도가 증가하며 시간 해상도가 감소한다. 반대의 경우 청크 크기를 줄이면 주파수 해상도가 감소하며 시간해상도가 증가하게 된다.

주파수 해상도가 증가하게되면 주파수 성분이 지속적으로 변하지 않는 신호에서 두각을 드러낸다. 반대로 시간 해상도가 증가하면 순간적인 변화 감지에서 두각을 드러낸다.

각 청크에 대해 FFT(Fast Fourier Transform)를 수행하여 주파수 영역으로 변환해준다. 이 과정으로 도메인 신호를 변환하여 주파수 성분을 분석할 수 있게 된다. 변환을 통해 얻어진 스펙트럼의 절대값을 계산한다. 그 다음, 주파수 범위(예: 40Hz, 80Hz, 120Hz, 180Hz, 300Hz)에 따라 각 범위 내에서 가장 강한 피크를 탐지한다. 이를 위한 방법으로 로그 스펙트

럼을 사용하고, 로그 스펙트럼을 통해 피크의 강도를 계산한 후 가장 강한 피크 주파수를 탐지한다. 주파수 범위는 일반적으로 인간의 청각 특성을 반영하는데, 인간의 귀는 특정 주파수 대역에서 더 민감하게 반응하는 특징이 있으므로, 이러한 대역을 중심으로 범위를 설정한다.

탐지된 피크 주파수를 기반으로 해시를 생성하는데 해시는 각 주파수 피크의 값을 특정 값으로 정규화한 후, 이를 결합하여 고유한 해시값을 생성하게 된다. 여기서 사용되는 정규화 값은 'FUZ_FACTOR'로 불리며, 이는 주파수 값의 정밀도를 줄여 근사치를 사용함으로써 잡음에 대한 강인함을 높이는 역할을 한다. 주파수 피크 값이 123일 때, 'FUZ_FACTOR'가 2라면, 이 값은 2의 배수인 122로 정규화된다. 이를 통해 근접한 주파수 값들을 같은 해시값으로 묶어 잡음과 미세한 변동에 대해 강인한 해시값을 생성할 수 있다.

Algorithm 1. Peak-based audio fingerprinting pseudocode

```
Definitions:
    RANGES = [40, 80, 120, 180, 300]
    FUZ_FACTOR = 2
    segment_length = 5 // Length of each audio segment in
    seconds

Function get_index(freq):
    for each val in RANGES:
        if freq 〈 val:
    return index of val
return last index in RANGES

Function hash_points(points):
    result = 0
    for each point in points reversed:
        result = result * 100000 + (point - (point %
        FUZ_FACTOR)) return result

Function get_fingerprints(audio, sample_rate):
    Define chunk_size, hop_size Initialize
    peaks list

    for each chunk in audio:
        Calculate spectrum
        Initialize points and highscores for frequency ranges
        Identify peak frequencies and update points list
        Calculate hash value from points and append to peaks
    return peaks

Main Matching Logic:
    queryBlock = get_fingerprints(query_audio,
    sample_rate)bestMusic = none
    minimumBER = 1,0
    for each fingerprint in queryBlock:
    indexNode = Lookup in database using fingerprint
    hash while indexNode is not null:
        Calculate BER with indexNode
        Update bestMusic if BER is new minimum and within
        threshold Move to next node in database chain
    return bestMusic
```

피크 기반 오디오 지문 생성은 계산이 효율적이고 잡음에 대한 저항성이 높다. 이는 특히 다양한 환경에서 안정적인 오디오 인식을 가능하게 한다. 다음 코드는 위에 설명한 피크 기반 오디오 지문 생성 코드를 예로 들은 주파수 범위에 따라 작성한 예시이다.

피크 기반 오디오 지문 생성의 장점은 잡음 저항성과 계산의 효율성이 있다. 이는 다양한 환경에서 안정적인 오디오 인식을 가능하게 해주며, 계산 과정 또한 단순하다. 특정 주파수 피크의 강도도 반영함으로써 잡음이 많은 환경에서도 오디오 신호를 정확하게 인식할 수 있다.

3.3 다중 시간 세그먼트 피크 기반 오디오 지문 생성

다양한 길이의 세그먼트를 통해 오디오의 시간적 특성을 더욱 상세하게 반영하기 위해, 다중 시간 세그먼트를 사용하는 방법을 제안한다. 이는 오디오 지문의 검색 정확도 또한 높여준다.

오디오 신호를 초 단위의 세그먼트로 나눈다. 그리고 각 세그먼트에 앞서 설명한 피크 기반 지문 생성 과정을 적용한다. 이러한 방식으로 생성된 지문은 데이터베이스에 저장되며, 검색할 때 가장 많이 검색된 콘텐츠가 선택된다. 이것은 다양한 길이의 오디오에서 오디오 신호의 특징을 추출하여, 잡음에 의한 변수를 줄이고, 정확도를 높이는 데에 도움이 된다.

주의할 점은 다중 시간 세그먼트를 사용할 때 생성 시간과 데이터의 크기를 생각해 샘플링 레이트를 낮추어야 한다. 본 연구에서는 샘플링 레이트를 5050Hz로 설정하여 크기를 줄이고 생성 시간을 단축시켰다.

4. 성능평가

4.1 전체 실험 환경

본 연구에서는 퍼블릭 도메인 영화 파일을 획득하여 해시 테이블 데이터베이스로 사용하였고, 해당 영화 파일의 무작위 시간대의 오디오 데이터를 저장하여 다중 시간 세그먼트 피크 기반 오디오 지문 검색에 사용하였다.

쿼리는 5초의 길이를 사용하였으며, 다중 시간 세그먼트 피크는 (1, 3, 5)초를 사용하였다. 잡음이 있는 상황을 가정하여 데이터에 5dB과 10dB의 잡음을 강제로 추가한 데이터와 안드로이드 환경에서 저장한 오디오 데이터를 사용하여 다양한 수준의 잡음에 대해 얼마나 내성이 있는지 테스트하였다.

Table 1. Fingerprint search accuracy

	5 sec Query	Multi Segments
clean	91.3%	87.8%
noise 5db	54.2%	68.7%
noise 10db	33.5%	50.3%
capture	78.4%	75.6%

4.2 성능평가

본 연구에서 설명한 피크 기반 오디오 지문 시스템과 다중 시간 세그먼트 피크 기반 오디오 지문 시스템의 결과를 측정하였다. 〈Table 1〉은 검색 정확도를 기록한 것이다.

실험 결과 잡음을 추가하지 않은 데이터와 안드로이드에서 직접 캡쳐한 데이터의 경우 5초 단일 쿼리로 생성된 피크 기반 오디오 지문 시스템의 성능이 더 뛰어났고, 노이즈가 추가될수록 정확도가 하락하였다. 하지만, 잡음을 주었을 경우에는 다중 세그먼트의 성능이 더 뛰어난 것도 확인할 수 있다.

〈Fig. 1〉은 샘플링 레이트와 쿼리 길이에 따른 지문 생성 시간을 보여주는 그래프이다. 그래프에서 알 수 있듯이 샘플링 레이트를 줄이며 다중 시간 세그먼트 기법을 사용하면, 단일 5초 쿼리를 사용했을 때와 지문 생성 시간이 유사하며, 실제 환경에서 발생하는 잡음과 같은 문제에는 성능이 더 좋은 것을 확인할 수 있었다.

5. 결 론

본 연구에서는 다중 시간 세그먼트를 사용하는 방법을 제안하였다. 다중 시간 세그먼트 피크 오디오 지문 시스템을 이용하여 콘텐츠를 식별하는 새로운 접근 방식을 개발하고, 실제 환경에서 테스트함으로써 가능성을 입증하였다.

본 연구의 핵심은 오디오 지문 기술의 정확성을 올리는 것이다. 딥러닝을 이용하여 기술의 정확도를 더 올리는 방법도 있겠지만, 근본적인 구조에 접근하여 특징적인 피크를 잡아내고, 다중 시간 세그먼트를 도입하여 기존의 방식대로 오디오 지문을 추출하여 콘텐츠를 검색하는 것보다 더 효율적이고 높은 정확도를 얻는 것을 목표로 하였다. 이렇게 높아진 검색 정확성으로 이전부터 연구되던 singing이나 humming으로 음악을 검색한다던가 OTT 콘텐츠의 특정 부분만으로 그 콘텐츠의 특정 시간대를 검색하는 것에 도움이 되어 그 응용 분야에 유용하게 쓰이기를 기대하고 있다. 또한 우리가 흔히 사용하는 특정 소프트웨어들이 아닌 제3의 하드웨어로 녹음, 녹화한 상황의 잡음까지 실험한 것이므로 꼭 음성파일을 가지고 있지 않더라도, 실시간으로 우리가 가진 모바일 기기로 녹음을 진행해도 성능이 나올 수 있는 것을 본 연구의 실험으로 입증했다.

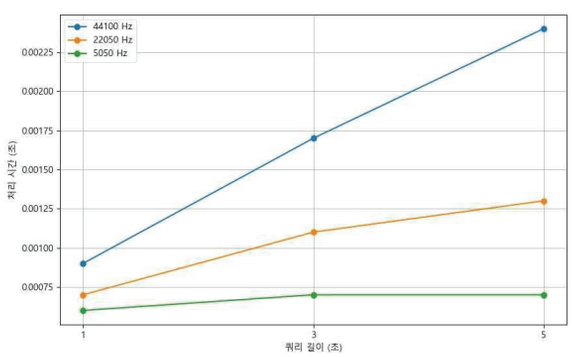

Fig. 1. Fingerprint creation time by sampling rate

추가로 본 논문은 향후 연구 방향에 대해 몇 가지를 제안한다. 먼저 오디오 저장 성능을 더욱 높이기 위해 알고리즘의 개선이 필요하다. 그 이후 딥러닝 기술을 도입하여 잡음에 더 강인한 시스템을 개발해야 한다. 마지막으로 데이터가 대규모로 들어올 경우, 데이터를 분산하여 처리할 수 있는 시스템의 도입도 고려하여야 한다. 이러한 연구 방향은 본 연구를 더욱 발전시키고, 실용성을 높여줄 것이다.

References

[1] M. Kim, M. Park and H. Kim, "Automatic Music Summarization Method by using the Bit ErrorRate of the Audio Fingerprint and a System thereo," *Journal of Korea Multimedia Society*, Vol.16, No.4, pp.453-463, 2013.

[2] D. Lee, M. Lim and J. H. Kim, "Music Recognition Using Audio Fingerprint: A Survey," *Phonetics and Speech Sciences*, Vol.4, No.1, pp.77-87, 2012.

[3] Shazam Entertainment Ltd., Shazam Announces Application for IPhone [Internet], http://www.shazam.com/music/web/pressrelease.html?nid=NEWS20080710151004.

[4] A. Wang, "An Industrial-Strength Audio Search Algorithm," in *Proceedings of ISMIR*, pp.7-13, 2003.

[5] S. H. Ryu and H. G. Kim, "Audio Fingerprinting Based on Constant Q Transform for TV Commercial Advertisement Identification", *The Journal of the Acoustical Society of Korea*, Vol.33, No.3, pp.210-215, 2014.

[6] D. H. Shin, J. S. Kim and C. W. Kim, "A Study for Utilization of Smart Device and Audio Fingerprinting Technologies to Help the Vision and Hearing Impaired People Consuming Broadcasting Contents Conveniently", *Journal of Information Technology and Architecture*, Vol.13, No.3, pp.457-466, 2016.

[7] M. Park and H. Kim, "An Audio Fingerprinting Scheme Robust to Real-Noise Environments", *Telecommunications Review*, Vol.16, No.3, pp.435-446, 2006.

[8] J. M. Kwon, I. J. Ko and D. S. Jang, "Search speed improved minimum audio fingerprinting using the difference of

Gaussian", *Journal of the Korea Society of Computer and Information*, Vol.14, No. 2, pp.75-87, 2009.

[9] M. Jia, T. Li and J. Wang, "Audio Fingerprint Extraction Based on Locally Linear Embedding for Audio Retrieval System", *Electronics*, Vol.9, No.9, 2020.

[10] K. Banuroopaa and D. S. Priyaaa, "MFCC based hybrid fingerprinting method for audio classification through LSTM", *International Journal of Nonlinear Analysis and Applications*, Vol.12, pp.2125-2136, 2021.

[11] S. I. Han, "Speech-Music Discrimination Using Deep Learning", *Journal of Korea Academia-Industrial cooperation Society*, Vol.22, No.10, pp.552-557, 2021.