JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 6, pp. 175-184, June 2026
https://doi.org/10.9708/jksci.2026.31.06.175

Music Sense: A Mobile Platform for Multimodal Music Perception for Individuals with Hearing Impairment

Sangyeop Kim*, Mirim Kim**, Jinwook Kim***, Jia Yoo****, Jeabum Lim*****, Pilsang Jung******, Won Joo Lee†

*Student, Dept. of Computer Science and Engineering, Soongsil University, Seoul, Korea
**Student, Dept. of Electronic and Information Engineering, Soongsil University, Seoul, Korea
***Student, Dept. of Data Science, Hanyang University, Seoul, Korea
****Student, Dept. of Mathematics and Statistics, Sejong University, Seoul, Korea
*****Master Student, Dept. of Math, Seoul National University, Seoul, Korea
******Student, Dept. of Energy Resources Engineering, Seoul National University, Seoul, Korea
†Professor, Dept. of Computer Science & Engineering, Inha Technical College, Incheon, Korea

[Abstract]

In this paper, we propose a Music Sense, a mobile platform–based application designed for individuals with hearing impairment and hard-of-hearing. The system adopts a client–server architecture in which computationally intensive AI-based audio analysis is performed on a cloud server, while the mobile client focuses on rendering synchronized visual and haptic feedback during playback. The server converts audio signals into time-series metadata through music source separation, pitch estimation, beat and onset detection, and phoneme-level lyric alignment. The generated metadata are normalized across multiple modalities and streamed in real time to mobile devices, where they are rendered as subtitles and visual representations of pitch and rhythm. Field evaluations conducted with children with hearing impairment demonstrate that the proposed application effectively supports rhythm reproduction, melody tracking, and the development of therapeutic music content. Furthermore, the proposed framework standardizes the conversion of music into reusable metadata and enables platform-level deployment via APIs and SDKs without modifying the original audio content, highlighting its scalability and practical applicability.

▸ **Key words**: Music accessibility, Multimodal interaction, Haptic feedback, Phoneme-level lyrics alignment, Mobile music system

• First Author: Sangyeop Kim, Corresponding Author: Won Joo Lee
*Sangyeop Kim (yeop@yeop.kr), Dept. of Computer Science and Engineering, Soongsil University
**Mirim Kim (kimmirim077@gmail.com), Dept. of Electronic and Information Engineering, Soongsil University
***Jinwook Kim (jwkim628@gmail.com), Dept. of Data Science, Hanyang University
****Jia Yoo (ujia0220@gmail.com), Dept. of Mathematics and Statistics, Sejong University
*****Jeabum Lim (max5972@naver.com), Dept. of Math, Seoul National University
******Pilsang Jung (snuisgood@snu.ac.kr), Dept. of Energy Resources Engineering, Seoul National University
†Won Joo Lee (wonjoo2@inhatc.ac.kr), Dept. of Computer Science & Engineering, Inha Technical College
• Received: 2026. 05. 26, Revised: 2026. 06. 12, Accepted: 2026. 06. 16.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 논문에서는 청각장애인(Hearing Impairment) 및 난청인(Hard-of-Hearing)을 위한 모바일 플랫폼 기반의 Music Sense 애플리케이션을 제안한다. 이 애플리케이션은 클라이언트-서버 구조로 계산 비용이 많이 소요되는 AI 분석을 서버에서 전처리하고, 클라이언트에서는 재생 시점에 맞춘 시각 및 촉각 피드백 렌더링 기능을 제공한다. 클라우드 서버에서는 오디오 신호를 시계열 메타데이터로 변환하기 위해 음원 분리, 피치 추정, beat/onset 분석, 음소 수준 가사 정렬 등을 수행한다. 이러한 시계열 메타데이터는 멀티-모달 정규화를 통해 실시간으로 모바일 디바이스에 전송되어 자막으로 제공되고, 음정과 박자로 렌더링한다. Music Sense 애플리케이션은 청각장애 아동을 대상으로 필드 테스트를 진행하였으며, 리듬 재현, 멜로디 추적, 치료용 콘텐츠 제작 등에 효과적으로 활용 가능하다는 결과를 얻었다. 또한, 본 논문의 연구 결과는 음악을 재사용할 수 있는 메타데이터로 변환하는 과정을 정규화하고, 원본 오디오의 수정 없이 API 및 SDK 기반의 플랫폼 단계로 배포 가능하다는 것이 장점이다.

▸ **주제어**: 음악 접근성, 멀티-모달 상호작용, 햅틱 피드백, 음소 수준 가사 정렬, 모바일 음악 시스템

I. Introduction

디지털 접근성은 부가 기능이 아닌 제품과 서비스의 기본 품질을 구성하는 핵심 요소로 인식되고 있다. 특히 유럽 접근성법(European Accessibility Act, EAA)은 접근성 요구사항을 제품 및 서비스 차원으로 명시하고 있으며, 2025년 6월 28일부터 신규 제품 및 서비스에 적용된다[1]. 이러한 변화는 음악 스트리밍, 교육, 치료, 엔터테인먼트 플랫폼이 청각장애인을 포함한 다양한 사용자의 감각 경험을 구조적으로 재설계해야 함을 의미한다. 그러나 현재의 음악 접근성 기능은 대체로 가사 자막, 간단한 비트 진동, 보조적 시각화 수준에 머물러 있다. 이러한 방식은 음악의 핵심 구성 요소인 리듬, 멜로디, 강세, 다이내믹, 가사 전달 타이밍을 통합적으로 제공하지 못한다. 특히 난청인이나 인공와우 사용자에게 음악은 단지 “들리는가”의 문제가 아니라, 시간적으로 정렬된 구조 정보가 얼마나 명확하게 전달되는가의 문제와 밀접하게 연결된다. 따라서 음악 접근성은 단순 번역이나 보조 디스플레이가 아니라, 오디오 구조를 정밀하게 분석하고 이를 촉각·시각 채널로 재구성하는 멀티-모달 정보 처리를 위한 접근이 필요하다.

기존 연구는 이러한 문제를 해결하기 위해 여러 유형의 촉각 음악 시스템이 제안되었으나 전용 장비 의존성, 범용성 부족, 가사 정렬 부재, 실시간 플랫폼 확장성 부족 등의 한계를 보였다. 최근에는 상용 모바일 시스템에서 음악과 햅틱을 결합한 접근성 기능이 등장하기 시작했으나[2][3], 시스템 내부 생성 방식이 폐쇄적이며, 보컬, 리듬, 가사, 구조적 전환을 분리하여 인지 가능한 형태로 제공하지 않는다. 따라서 AI 기반의 정밀 분석으로 플랫폼에 적용 가능한 메타데이터 형태로 변환함으로써 범용적으로 모바일 디바이스에서 동기화 재생할 수 있는 서비스가 필요하다.

본 논문에서는 청각장애인과 난청인을 위한 AI 기반 멀티-모달 음악 서비스로 Music Sense 애플리케이션을 제안한다. 2장에서는 촉각 음악, 오디오-촉각 지각, AI 기반 접근성 시스템 관련 선행 연구를 분석한다. 3장에서는 Music Sense 애플리케이션을 설계하고, 4장에서는 Music Sense 애플리케이션 구현 방법에 대하여 설명한다. 5장에서는 결론과 확장성을 제시한다.

II. Preliminaries

1. Tactile Music System for the Hearing Impaired

청각장애인을 위한 촉각 음악 연구는 고정형 설치 장치, 착용형 장치, 모바일 장치 분야가 있다. S. Nanayakkara[4]은 Haptic Chair를 통해 의자 진동과 시각 디스플레이를 결합하여 청각장애인 및 난청인이 저역 리듬과 음악 구조를 감지하도록 설계하였다. 이 시스템은 청각장애인을 위한 몰입형 음악 경험의 가능성을 보여주었으나, 대형 설치물에 의존하기 때문에 가정, 학교, 복지관, 이동 환경에서 범용적으로 사용하기에는 한계가 있다. 후속 연구에서는 Haptic Chair를 음성 훈련 보조 도구로 확장하여, 청각장애인 20명을 대상으로 발화 훈련과 촉각

피드백 결합의 치료적 적용 가능성을 탐색하였다. 이는 촉각 시스템이 감상형 인터페이스를 넘어 재활 및 훈련으로 사용할 수 있으며, 소규모 파일럿 수준이나 후속 임상 연구의 출발점이다. 하지만 전용 하드웨어와 특정 사용 시나리오에 강하게 결합되어 있어, 대규모 콘텐츠 플랫폼이나 일상형 모바일 서비스로 바로 연결되기는 어렵다.

M. Karam[5, 6]은 음성 및 음악의 정서적 요소를 촉각적으로 전달하는 크로스 모달 오디오-촉각 디스플레이를 설계하였다. 이 연구는 청각 정보를 촉각 패턴으로 치환하는 체계적 접근을 보여주었지만, 장치 설계 자체가 연구 핵심이기 때문에 일반 스마트폰 환경으로의 이식성은 제한적이었다.

저역과 고역을 분리하여 모바일 장치에서 실시간 햅틱 음악을 재생하는 모바일 기반의 real-time dual-band haptic music player[7]는 기존보다 정밀도와 선호도 측면의 개선 가능성을 보였다. 그러나 이 방식은 기본적으로 대역 기반 햅틱 재생에 초점을 두고 있어, 가창의 음정 윤곽이나 가사 전달 타이밍까지 포괄하는 구조적 음악 접근성 시스템으로 확장되지는 않았다.

ETRI의 Tactile Tone System은 보다 음정 훈련 문제를 다룬다. 9개의 진동 모터와 36개 음높이(C3-B5)를 매핑한 웨어러블 장치를 제안하고, 2명의 인공와우 사용자에게 15시간의 보컬 훈련을 제공하여 음정 정확도 향상을 보였다[8]. 이 연구는 피치-진동 매핑의 훈련 효과를 보여준다는 점에서 중요하지만, 주된 목표가 노래 훈련이며, 전체 곡 감상 경험, 가사 자동화,리듬 멜로디 시각 정보의 동시 통합을 제공하지는 않는다.

2. Audio-tactile Crossmodal Perception and Synchronization

멀티-모달 시스템에서 가장 중요한 기술적 문제 중 하나는 시간 정렬이다. M. Zampini[9]는 오디오-촉각 temporal order judgment 실험을 통해 숙련 관찰자의 경우 약 40ms, 비숙련 관찰자의 경우 약 80ms 수준에서 자극 순서 판단 성능이 달라질 수 있음을 보고하였다. 이는 40ms를 보편적 절대 기준으로 해석해야 한다는 뜻은 아니지만, 정밀 동기화가 실제 지각 품질에 직접적인 영향을 미친다는 점을 보여준다.

C. M. Karns[10]는 선천성 청각장애인의 1차 청각 피질에서 시각 및 촉각 입력이 재조직화될 수 있음을 fMRI 기반으로 제시하였다. 이는 청각장애인을 위한 멀티-모달 인터페이스가 단순한 대체 출력 채널이 아니라, 실제 감각 처리 구조와 결합될 수 있음을 의미한다. 다시 말해, 촉각과 시각은 청각 정보의 보조가 아니라 청각 경험의 재구성 채널로 이해될 필요가 있다.

3. AI-based Music Accessibility and Inclusive Instrument Interface

E. Frid[11]는 포용적 음악 실천 맥락에서 활용된 83개의 Accessible Digital Musical Instruments(이하 ADMI)에 대한 리뷰에서 사용자 맞춤성, 다감각 피드백, 참여적 설계를 포용형 음악 인터페이스의 핵심 원칙으로 제시함으로써 기존 악기에 단순히 장애 수용 기능을 부가하는 수준을 넘어, 감각 채널과 상호작용 방식을 초기 설계 단계부터 재구성해야 한다고 주장하였다. 본 연구는 이러한 포용형 설계 원칙을 연주 중심 인터페이스가 아닌 감상 중심 모바일 서비스로 확장한 사례로 이해할 수 있다.

4. Analysis of Previous Research

선행 연구 분석 결과 종합하면 다음과 같은 연구 공백이 확인된다. 첫째, 청각장애인을 위한 촉각 음악 시스템의 상당수는 햅틱 의자, 고정형 설치 장치, 웨어러블 글러브 등 특수 하드웨어에 의존하고 있어 일상적 활용성과 보급성에 한계가 있다[4-6, 8, 12]. 둘째, 기존 연구는 리듬 전달, 음정 훈련, 정서적 촉각 치환 등 특정 기능에 초점을 둔 경우가 많으며, 가사-음소 수준 정렬과 리듬 멜로디의 동시 매핑을 하나의 파이프라인으로 통합한 사례는 많지 않다[4-8, 12]. 셋째, 대부분의 연구가 장치 단위 실험 또는 사용자 반응 검증에 머물러 있어, 모바일 서비스 및 플랫폼 SDK 수준의 확장 구조를 명시적으로 설계한 사례가 제한적이다[4-8, 12]. 넷째, 음정 훈련이나 재활 보조 시스템과 음악 감상 시스템이 분절적으로 발전해 왔기 때문에, 치료 교육 상용 서비스 간 연결 구조가 상대적으로 약하다[8, 9, 12].

이러한 한계는 실제 음악 소비 환경에서 더욱 두드러진다. 음악 서비스 시장에서는 이미 존재하는 대규모 음원 카탈로그를 얼마나 정밀하게 분석하고, 이를 다양한 사용자군이 이해 가능한 멀티-모달 형식으로 변환하여 제공할 수 있는지가 핵심 과제로 부상하고 있다. 그러나 기존 연구는 대부분 특정 실험 장치의 효과 검증에 집중되어 있어, 곡 단위 분석 자동화, 메타데이터화, 모바일 렌더링, 서비스 재사용성까지 포함한 전체 파이프라인 수준의 접근은 부족하다. 따라서 음악 접근성은 더 이상 단일 장치의 촉각 출력 문제가 아니라, 음악 구조 분석, 시간 정렬, 멀티모달 렌더링, 플랫폼 배포를 포괄하는 시스템 문제이다.

Table 1. Comparison of prior studies and the proposed system

System	Target users	Platform / hardware	Main features	Service scalability	Main limitation
Nanayakkara et al. (2009) [4]	DHH users	Haptic chair + visual display	Low-freq rhythm, tactile music	Low	Fixed, low portability
Nanayakkara et al. (2012) [12]	Deaf users	Haptic chair	Tactile-based speech training	Low	Specialized, training-only
Karam et al. (2008, 2009) [5][6]	General / research users	Cross modal audio-tactile display	Audio→tactile mapping, emotion cues	Low	Ambient setup, hard scaling
Shin et al. (2020) [8]	Cochlear implant users	Wearable tactile glove	Pitch→vibration for vocal training	Low	Wearable-dependent, no full-song support
MUSIC SENSE (proposed)	DHH, therapy/edu, platform ops	General mobile device	Stem sep., phoneme alignment, pitch/rhythm, tactile-visual	High	PoC, iOS-only

표 1에서 확인할 수 있듯이, MUSIC SENSE의 차별점은 단순히 모바일 환경에서 작동한다는 점에만 있지 않다. 제안 시스템은 범용 모바일 디바이스를 실행 환경으로 삼으면서도, 음원 분리[13][14], 음정 추정[15], 리듬 검출[16][17], 가사 정렬[18]을 연계하여 리듬·멜로디·가사 강조 정보를 하나의 메타데이터 파이프라인으로 통합한다는 점에서 기존 연구와 구분된다. 특히 음소 수준 정렬과 모바일 햅틱 재생을 함께 다룸으로써, 단순한 비트 진동이나 대역 기반 촉각 재생보다 시간 정합성이 높은 멀티-모달 경험을 제공할 수 있다. 또한 분석 결과를 오디오 자체가 아닌 메타데이터 기반 전달 구조로 설계함으로써, 서비스 사업자가 기존 재생 엔진을 유지한 채 접근성 기능을 결합할 수 있다. 이는 단일 연구 장치의 효과를 입증하는 수준을 넘어, 치료·교육·상용 플랫폼에 공통으로 적용 가능한 구조를 제안한다. 즉, 본 연구의 차별성은 새로운 촉각 장치를 제안하는 데 있는 것이 아니라, 기존 음악 콘텐츠를 대규모로 분석·재가공하여 범용 모바일 환경에서 재사용 가능한 접근성 서비스로 전환할 수 있다.

III. The Design of Music Sense Application

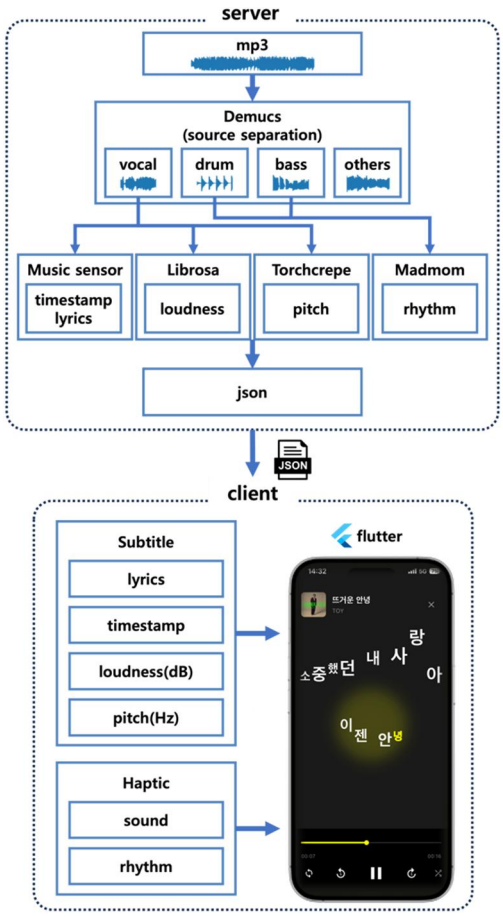

Fig. 1. Music Sense Application Architecture

본 논문에서 제안하는 Music Sense 애플리케이션은 그림 1과 같이 클라이언트-서버로 구성된다. Demucs, Music sensor, Librosa, Torchcrepe, Madmom은 클라우드 서버에서 실행되는 모듈이고, Subtitle, Haptic은 클라이언트인 스마트폰에서 실행되는 모듈이다. 이러한 클라이언트-서버 구조의 장점은 계산 비용이 많이 소요되는 AI 분석을 서버에서 전처리하고, 클라이언트는 재생 시점에 맞춘 시각 및 촉각 피드백 렌더링 기능을 제공한다. AI 분석 결과는 원본 음원 자체가 아닌 시간축 기반 멀티-모달 메타데이터이다. 따라서 서비스 사업자는 음원을 재배포하지 않고도 접근성 레이어를 기존 콘텐츠에 부가할 수 있기 때문에 저작권 위험을 줄이면서 B2B SDK 또는 API 형태의 배포를 가능하다.

IV. Implementation and Experimental Results of MUSIC SENSE Application

1. Implementation of Asynchronous Analysis Pipeline

MUSIC SENSE의 서버 측 구현은 입력 음원을 시간 축 기반 멀티모달 메타데이터로 변환하는 비동기 분석 파이프라인으로 구성된다. 서버는 먼저 업로드된 음원의 포맷, 길이, 샘플링 조건을 확인한 뒤, 이후 단계에서 사용될 보컬, 리듬, 가사 정렬 정보를 추출한다. 즉, 계산 비용이 큰 분석 과정은 서버에서 수행하도록 구현하고, 모바일 클라이언트에서는 분석 결과를 재생 시점에 맞추어 렌더링하도록 구현한다.

제안 파이프라인의 첫 단계는 혼합 음악 신호를 보컬, 드럼, 베이스 및 기타 반주 스템으로 분리하는 것이다. 원신호에는 다양한 악기와 보컬이 동시에 중첩되어 존재하므로, 이를 그대로 입력하면 후속 피치 추정, 리듬 검출, 가사 정렬 과정에서 상호 간섭이 발생할 수 있다. 이러한 문제를 줄이기 위해 Demucs 계열의 딥러닝 기반 음악 소스 분리 모델을 사용하였다[13][14]. Demucs 기반 분리기의 도입 목적은 분리 성능 자체의 비교가 아니라, 후속 분석 모듈의 안정성을 높이는 데 있다. 구체적으로 분리된 보컬 스템은 CREPE 기반 피치 추정과 가사 정렬 입력으로 사용하고[15][18], 드럼 및 리듬 중심 스템은 madmom 기반 beat/onset 분석 입력으로 사용한다[16][17]. 이를 통해 보컬 피치 곡선은 반주 성분의 영향을 덜 받게 되고, 리듬 이벤트 역시 화성 악기나 잔향에 의한 오검출을 줄일 수 있다.

가사 정렬은 단어 수준 timestamp만으로는 충분하지 않다. 실제 노래는 장음, 빠른 발화, 멜리스마, 박자 내 음절 분할 등으로 인해 단어 내부의 시간 구조가 크게 달라진다. 따라서 본 연구는 Schulze-Forster의 phoneme level lyrics alignment 연구를 바탕으로, 음절·음소 수준 정렬을 지원하는 모델을 채택한다[18]. 음소 정렬 오차는 식 (1)과 같이 정의한다.

$$E_{phoneme} = \frac{1}{N}\sum_{i=1}^{N}\left|t_i^{pred} - t_i^{ref}\right| \quad (1)$$

위의 식 (1)에서 N은 정렬된 음소수, t_i^{pred}는 모델이 예측한 i번째 음소의 시작 시각, t_i^{ref}는 기준 시각을 의미한다. 본 논문에서는 오디오-촉각 temporal order judgment 문헌에서 보고된 숙련 관찰자 수준의 정밀도를 참고하여, 수십 ms 수준의 정렬 오차를 설계 목표로 설정한다[9]. 특히, 한국어 가창의 경우 초성-중성-종성으로 구성되는 음절 구조와 장음 중심의 발화 특성 때문에 영어 중심 음소 정렬 모델을 그대로 적용하기 어렵다. 따라서 한국어 가사를 우선 음절 단위로 분절한 뒤, 이를 정렬 입력의 기본 단위로 사용한다. 또한 늘임 음과 반복 음절이 나타나는 구간에서는 음절 체류시간을 별도로 보정하여, 가창 구간의 비선형 시간 구조를 보다 안정적으로 반영하도록 한다. 이러한 방식은 한국어 singing alignment 데이터 셋이 충분하지 않은 상황에서 적용 가능한 실용적 절충안에 해당하며, 향후 한국어 가창 특화 정렬 데이터셋 구축을 통해 추가 개선이 가능하다.

또한 본 논문에서는 단일 정렬 모델에 전적으로 의존하지 않고, 그림 2와 같이 WhisperX와 Lyrics-Aligner를 상보적으로 결합하는 앙상블 전략을 채택한다.

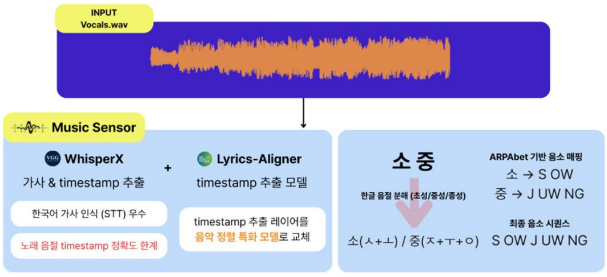

Fig. 2. Detail of WhisperX and Lyrics-Aligner Layer

그림 2에서 WhisperX는 VAD와 forced phoneme alignment를 이용해 장문 오디오에서 단어 수준의 time-accurate timestamp를 제공하는 데 강점을 가지며[19], Lyrics-Aligner는 노래 전용 음향-음소 매칭과 DTW 경로 탐색을 통해 음절 수준 정렬에 강점을 가진다[18]. 따라서 본 논문에서는 WhisperX가 제공하는 단어 경계 후보를 coarse anchor로 사용하고, Lyrics-Aligner의 DTW 기반 정렬 경로로 세부 음절 위치를 보정하는 2단계 구조를 사용한다. 이를 통해 멜리스마 구간에서는 WhisperX의 오정렬을 DTW 경로로 보정하고, 반대로 빠른 발화 구간에서는 WhisperX의 경계 정보로 DTW 탐색 범위를 안정화한다.

보컬 멜로디 추출에는 CREPE를 적용한다[15]. CREPE는 파형 입력으로부터 단음(monophonic) 피치를 직접 추정하는 CNN 기반 모델이며, 본 논문에서는 Demucs로 분리한 보컬 스템에 적용하여 단일 멜로디 윤곽을 안정적으로 확보한다. 이는 폴리포닉 혼합 신호에 CREPE를 직접

적용하는 경우보다 보컬 중심 피치 곡선을 더 명확하게 추출할 수 있도록 한다. 리듬 특징은 madmom 기반 beat/onset 분석으로 추출한다[16][17]. Onset strength 계산 단계에서는 SuperFlux 계열의 vibrato suppression 개념을 반영하며, 이를 개념적으로 표현하면 식 (2)와 같다.

$$Flux_{SF}(t) = \sum_{k} \max(S(k,t) - \max_{u=t-\mu}^{t-1} S(k,u), 0) \tag{2}$$

식 (2)에서 $S(k,t)$는 주파수 빈 k에서의 시간 t의 스펙트럼 크기이며, μ는 최대 필터 구간 길이를 의미한다. 즉, 인접 프레임 간 단순 차분이 아니라 직전 구간의 최대값과 비교함으로써 vibrato나 미세한 음고 흔들림으로 인한 오검출을 줄인다[16]. 또한 MUSIC SENSE 앱은 madmom의 RNNBeatProcessor와 DBNBeatTrackingProcessor를 체인 방식으로 결합한다[17]. RNNBeatProcessor가 각 프레임의 beat activation $P_{RNN}(t)$를 산출하면, DBNBeatTrackingProcessor가 템포 사전확률과 시계열 전이 제약을 반영하여 최종 비트 위치를 결정한다. 이를 개념식으로 표현하면 식 (3)과 같다.

$$t^{*} = \arg\max_{t} [P_{RNN}(t) \cdot P_{DBN}(t|tempo)] \tag{3}$$

이 구조는 단순 onset peak picking보다 음악적 맥락을 반영한 beat 추적이 가능하다는 것이 장점이다. 서버에서 추출된 보컬 피치, 리듬 강도, 에너지, 가사 강조도는 단위와 분포가 서로 다르므로, 동일한 출력 채널에서 직접 매핑할 수 없다. 따라서 각 특징을 0-1 범위로 정규화한 후 모바일 렌더링에 사용한다. 정규화 과정은 식 (4)와 같이 정의한다.

$$f(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

식 (4)에서 x는 원시 특징값, $x_{\min}$과 $x_{\max}$는 곡 또는 세그먼트 내 최소·최대값이다. MUSIC SENSE 앱애서는 곡 단위 정규화를 기본으로 하되, 급격한 동적 범위 변화가 있는 경우 구간 단위 재정규화를 추가 적용한다.

햅틱 sharpness는 리듬의 선명도와 멜로디 변화량을 동시에 반영하도록 식 (5)와 같이 정의한다.

$$Sharp_{final}(t) = \alpha \cdot \widehat{Flux}(t) + \beta \cdot |\frac{d\hat{f}_0(t)}{dt}| \tag{5}$$

식 (5)에서 $\widehat{Flux}(t)$는 정규화된 리듬 온셋 강도, $\hat{f}_0(t)$는 정규화된 보컬 피치 곡선, α와 β는 각각 리듬 및 멜로디 변화에 대한 가중치이다. MUSIC SENSE 앱에서는 α = 0.6, β = 0.4를 기본값으로 설정한다. 이는 리듬 cue를 햅틱 출력의 우선 정보로 두되, 멜로디 윤곽 변화가 완전히 소실되지 않도록 한 경험적 설정값이다. 특히 보컬 채널과 드럼 채널을 동시에 출력할 때 sharpness를 단순 합산하지 않고, 각 채널의 가중치를 두어 최종 sharpness를 구성한다.

서버 구현 환경은 RTX 4090 laptop GPU 기준으로 구성되었고, 실험 음원 1곡당 약 8~15초 수준의 분석 시간이 소요된다. 분석이 완료된 결과는 가사 timestamp, 피치 곡선, beat/onset, intensity, sharpness를 포함하는 JSON 형태의 메타데이터로 직렬화된다. 한 번 분석된 곡의 결과는 캐시에 저장되어 재사용할 수 있으며, 이후 클라이언트는 원본 음원을 직접 수정하지 않고도 해당 메타데이터를 기반으로 시각·촉각 피드백을 동기화하여 재생할 수 있다.

2. Implementation of Client Rendering Process

클라이언트 측 구현은 서버에서 생성된 시계열 메타데이터를 실제 재생 타임라인에 맞추어 시각 및 촉각 정보로 변환하는 과정이다. 본 논문에서 모바일 클라이언트는 iOS 환경에서 구현되었으며, AVFoundation 기반 오디오 재생 타임라인과 Core Haptics를 연동하여 자막, 시각화, 햅틱 피드백을 동기화한다[3].

시각 채널에서는 음 높이 변화에 따라 가사 위치, 글자 굵기, 오브젝트 크기 또는 색상값을 변화시켜 멜로디의 상·하행을 표현한다. 리듬 이벤트는 박자 경계, 하이라이트, 애니메이션 점멸, 구간 전환으로 표시되며, 가사 채널은 현재 재생 중인 음절 또는 음소를 기준으로 강조된다. 긴 음절의 경우 하이라이트 지속 시간을 늘려, 음가와 시각 체류 시간이 어긋나지 않도록 구현한다.

촉각 채널에서는 진동 모듈의 intensity와 Sharpness를 이용하여 에너지와 구조 변화를 제어한다[3]. MUSIC SENSE 앱은 10ms 단위 진동 profile을 생성하여 보컬 채널과 박자 채널을 분리한 JSON 형태로 관리한다. 실제 구현에서는 vocals.json과 drums.json의 두 채널을 생성하고, 이를 진동 모듈로 전달하여 기기 진동을 제어한다. 이때, 보컬 추종 채널과 박자 채널을 동시에 출력할 경우,

sharpness 값을 단순 합산하지 않고 보컬과 드럼의 가중치를 두어 최종 sharpness를 구성하도록 구현한다.

그림 3의 앱 화면에서는 가사·시각화 인터페이스, 자동 스크롤, 밀리초 단위 동기화, 리듬 기반 햅틱이 출력된다. 사용자는 음악 재생 중 현재 가사 위치와 음절 강조를 시각적으로 확인할 수 있으며, 동시에 리듬과 멜로디 변화에 대응하는 촉각 피드백을 경험할 수 있다. 이러한 구조는 청각장애 사용자에게 음악의 리듬, 멜로디, 가사 타이밍을 별도의 감각 채널로 재구성하여 제공한다는 점에서 단순 자막 또는 단순 비트 진동과 구분된다.

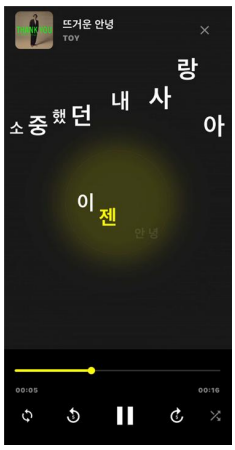

Fig. 3. App Demo

이와 같은 시각 자막 설계는 청각장애 사용자를 위한 발화 스타일 시각화 캡션 연구의 흐름과도 부합한다. Ahn 등은 punctuation, paint-on, color, boldness 등을 활용해 발화 스타일을 시각화한 캡션 시스템을 제안하였고, 26명 청각장애 참가자를 대상으로 한 비교 연구를 통해 이러한 시각 요소가 일반 캡션보다 발화 스타일 전달과 몰입감 측면에서 유의미한 가능성을 보인다[20]. MUSIC SENSE 앱은 이를 음악 감상 맥락으로 확장하여 보컬의 강세, 음절 체류 시간, 리듬 경계를 자막 스타일과 촉각 출력에 동시에 반영한다.

3. Experimental Results

본 논문의 실험은 청각장애 아동 3명을 대상으로 한 탐색적 적용 사례이다. 대상자는 인공와우 착용 아동, 보청기 착용 아동, 청신경병과 인공와우를 함께 가진 아동으로 구성한다. 검증 방식은 Gordon의 PMMA를 참고한 리듬 중심 평가 구조를 사용한다[21]. 사전 검사는 음악치료사의 피아노 반주에 맞추어 노래하도록 구성하고, 사후 검사는 동일한 날짜에 MUSIC SENSE 앱의 진동 및 시각 단서를 활용하여 노래하도록 구성한다. 대표곡은 “우리의 소원은 통일"을 사용하였으며, 주요 기록 지표는 리듬 정확도, 멜로디 정확도, 리듬 오류반응으로 설정한다. 실험 설계의 개요는 Table 2와 같다.

Table 2. Exploratory application design for children

Item	Description
Participants	3 hearing-impaired children
Hearing conditions	Cochlear implant, hearing aid, auditory neuropathy + cochlear implant
Evaluation framework	PMMA-based rhythm-focused assessment [21]
Pre-condition	Singing with therapist’s piano accompaniment
Post-condition	Singing with MUSIC SENSE using tactile and visual cues
Representative song	“우리의 소원은 통일”
Main recorded measures	Rhythm accuracy, melody accuracy, rhythm error response

탐색적 적용 결과, 세 아동 모두에서 사전 검사 대비 사후 검사의 리듬·멜로디 성과가 개선되었다. 리듬 정확도는 청신경병·인공와우 아동이 7/16에서 10/16, 인공 와우아동이 7/16에서 13/16, 보청기 아동이 11/16에서 14/16으로 증가하였다. 멜로디 정확도는 각각 33%에서 55%, 31%에서 50%, 50%에서 62.55%로 향상되었다. 리듬 오류반응은 각각 57%에서 33%, 57%에서 19%, 32%에서 13%로 감소하였다. 이 결과는 단일 세션 내 비교에서 MUSIC SENSE의 시각·촉각 단서가 리듬 및 멜로디 인지에 긍정적 보조 역할을 할 가능성을 보여준다.

결과 요약은 Table 3과 같다.

Table 3. Pre/post results of the exploratory child application

	Auditory neuropathy + CI child	CI child	Hearing-aid child
Rhythm (pre)	7/16	7/16	11/16
Rhythm (post)	10/16	13/16	14/16
Melody (pre)	33%	31%	50%
Melody (post)	55%	50%	62.55%
Rhythm error (pre)	57%	57%	32%
Rhythm error (post)	33%	19%	13%

음악치료사의 인터뷰에 따르면 기존 치료실에서는 스케치북과 스티커를 이용해 시각 자료를 수작업으로 제작하고 있으며, 음원당 평균 2시간 이상이 소요되는 한계가 있다. 반면 MUSIC SENSE는 생성된 메타데이터를 바탕으로 악보를 자동 생성하고, 스케일에 따라 색상을 다르게 부여하며, 박자에 따라 마디를 분리하여 아동의 눈높이에 맞춘 시각 자료를 생성할 수 있는 구조를 갖는다. 또한 현재 청음복지관의 “우리의 소원은 통일" 음악 교육에 실제 적용 중이다. 이러한 결과는 MUSIC SENSE가 사용자 경험 향상뿐 아니라 치료사의 콘텐츠 준비 부담을 줄이는 측면에서도 의미가 있다. 다만 본 연구에서는 엄격한 통제 실험이나 대규모 사용자 연구가 아니라 아동 3명을 대상으로 한 탐색적 적용 사례라는 점에서 한계를 가진다. 따라서 본 결과를 일반적 효과로 확대해석할 수는 없다. 또한 현재 검증은 특정 곡과 특정 치료 환경에 한정되어 있으며, 장르 다양성, 장기 학습 효과, 기기 차이에 따른 햅틱 편차, 개인별 촉각 민감도 차이 등은 별도의 후속 연구가 필요하다. 실험 결과는 MUSIC SENSE 앱이 청각장애 아동의 음악치료와 교육에 활용할 수 있다는 것을 검증한 것이다.

V. Conclusions

본 논문에서는 AI 기반 멀티모달 음악 접근성 파이프라인과 MUSIC SENSE 앱을 구현하였다. MUSIC SENSE 앱은 음원 분리, 음정 추정, beat/onset 분석, 음절·음소 수준 가사 정렬, 멀티모달 정규화, 시각·촉각 렌더링을 하나의 파이프라인으로 통합함으로써, 청각장애 사용자가 음악의 리듬, 멜로디, 가사 타이밍을 더 구조적으로 인지할 수 있도록 설계하였다. MUSIC SENSE 앱은 원본 음원을 재가공하거나 별도의 접근성 콘텐츠를 수작업 제작하지 않더라도, 곡의 구조 정보를 시간 정렬된 멀티모달 메타데이터로 변환해 제공하고 이를 클라이언트가 렌더링하는 방식으로 구현하였다. 이 구조는 서버측에서 고비용 분석을 수행하고, 클라이언트는 경량 메타데이터에 기반해 촉각·시각을 동기 재생하도록 구현하였다.

MUSIC SENSE 앱의 실험은 청음복지관과 협력하여 청각장애 아동 3명을 대상으로 리듬 및 멜로디 수행의 향상과 리듬 오류 반응 감소를 관찰하였다. 실험 결과는 MUSIC SENSE 앱이 청각장애 아동의 음악치료와 교육에 효과적임을 보인다.

향후 연구는 표본 확장 및 장기 반복 세션을 통한 효과 지속성 검증, 사용자별 촉각 민감도 및 기기 차이를 반영하는 캘리브레이션/개인화, 한국어 가창 정렬 고도화를 위한 데이터 축적 및 모델 개선을 중심으로 진행할 필요가 있다.

ACKNOWLEDGEMENT

This work was supported by the SK Telecom FLY AI Challenger Program, under the 2025 K-Digital Training Program jointly operated by the Ministry of Employment and Labor and the Korea Skills Quality Authority. The authors also gratefully acknowledge the Cheong-Eum Welfare Center for providing access to participants and supporting the exploratory evaluation of the proposed system.

REFERENCES

[1] European Parliament and Council of the European Union, “Directive (EU) 2019/882 of 17 April 2019 on the accessibility requirements for products and services,” Official Journal of the European Union, Apr. 2019.

[2] Apple Inc., “Use Music Haptics on your iPhone,” Apple Support, 2024.

[3] Apple Inc., “Core Haptics,” Apple Developer Documentation, 2024.

[4] S. C. Nanayakkara, E. A. Taylor, L. Wyse, and S. H. Ong, “An Enhanced Musical Experience for the Deaf: Design and Evaluation of a Music Display and a Haptic Chair,” in Proceedings SIGCHI Conference. Human Factors in Computing Systems, pp. 337-346, 2009. DOI: 10.1145/1518701.1518756

[5] M. Karam, F. Russo, C. Branje, E. Price, and D. Fels, “Towards a Model Human Cochlea: Sensory Substitution for Crossmodal Audio-Tactile Displays,” in Proceedings Graphics Interface, pp. 267-274, 2008.

[6] M. Karam, F. A. Russo, and D. I. Fels, “Designing the Model Human Cochlea: An Ambient Crossmodal Audio-Tactile Display,” IEEE Transactions on Haptics, Vol. 2, No. 3, pp. 160-169, July-Sept. 2009. DOI: 10.1109/TOH.2009.32

[7] I. Hwang, H. Lee, and S. Choi, “Real-time dual-band haptic music player for mobile devices,” IEEE Transactions on Haptics, Vol. 6, No. 3, July-Sept. 2013. DOI: 10.1109/TOH.2013.7

[8] S. Shin, C. Oh, and H. Shin, “Tactile Tone System: A Wearable Device to Assist Accuracy of Vocal Pitch in Cochlear Implant Users,” in Proceedings 22nd Int’l ACM SIGACCESS Conf. on

Computers and Accessibility, Article 61, 2020. DOI: 10.1145/3373625.3418008

[9] Massimiliano Zampini, Timothy Brown, David I. Shore, Angelo Maravita, Brigitte Röder, Charles Spence, "Audiotactile temporal order judgments," Acta Psychologica, Vol. 118, No. 3, pp. 277-291, Mar. 2005.

[10] C. M. Karns, M. W. Dow, and H. J. Neville, "Altered Cross-Modal Processing in the Primary Auditory Cortex of Congenitally Deaf Adults," Journal of Neuroscience, Vol. 32, No. 28, pp. 9626-9638, July 2012.

[11] E. Frid, "Accessible Digital Musical Instruments—A Review of Musical Interfaces in Inclusive Music Practice," Multimodal Technologies and Interaction, Vol. 3, No. 3, pp. 57, July 2019. DOI: 10.3390/mti3030057

[12] S. Nanayakkara, L. Wyse, and E. A. Taylor, "The Haptic Chair as a Speech Training Aid for the Deaf," in Proceedings 24th Australian Computer-Human Interaction Conference, pp. 405-410, 2012. DOI: 10.1145/2414536.2414600

[13] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music Source Separation in the Waveform Domain," arXiv:1911.13254, 2019.

[14] A. Défossez, "Hybrid Spectrogram and Waveform Source Separation," arXiv:2111.03600, 2021.

[15] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A Convolutional Representation for Pitch Estimation," in Proceedings IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 161-165, 2018. DOI: 10.1109/ICASSP.2018.8461329

[16] S. Böck and G. Widmer, "Maximum Filter Vibrato Suppression for Onset Detection," in Proceedings 16th Int'l Conference on Digital Audio Effects (DAFx), 2013.

[17] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: a new Python audio and music signal processing library," arXiv:1605.07008, 2016.

[18] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau, "Phoneme Level Lyrics Alignment and Text-Informed Singing Voice Separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 29, pp. 2382-2395, June 2021. DOI: 10.1109/TASLP.2021.3091817

[19] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," INTERSPEECH 2023 / arXiv:2303.00747, 2023.

[20] S. Ahn, J. Kim, C. Shin, and J.-H. Hong, "Visualizing speech styles in captions for deaf and hard-of-hearing viewers," International Journal of Human-Computer Studies, Vol. 194, 103386, Feb. 2025.

[21] E. E. Gordon, Primary Measures of Music Audiation. Chicago, IL, USA: GIA Publications, 1979.

Authors

Sangyeop Kim is currently pursuing a B.S. degree in the Department of Computer Science and Engineering at Soongsil University, Seoul, South Korea, where he is expected to graduate in August 2026. He has professional experience as a full-stack developer at a startup, where he was involved in web platform development using Django and Vue.js. He also completed an AI and Data Science summer program at the University of Southern California, Los Angeles, CA, USA, in 2025. He is interested in music information retrieval, audio processing, AI pipeline engineering, full-stack development, and cloud computing.

Mirim Kim will receive the B.S. degree in Electronic and Information Engineering from Soongsil University, Korea, in February 2027. Her research interests include machine learning, signal processing, and electromagnetic wave propagation modeling.

Jinwook Kim will receive the B.S. degree in Data Science from Hanyang University, Korea, in February 2027. His research interests include machine learning, statistical data analysis, System architecture and recommender system.

Jia Yoo received the B.S. degrees in Mathematics and Statistics and Data Science (double major) from Sejong University, Korea, in 2026. Her research interests include machine learning, deep learning, and multimodal AI systems.

Jeabum Lim received the B.A. degree in Physics from the University of Oxford, UK, in 2023, and will receive the M.S. degree in Mathematical Sciences from Seoul National University, Korea, in 2028. His research interests include computational physics, Bayesian statistics, and quantitative modeling.

Pilsang Jung is currently pursuing the B.S. degree in Energy Resources Engineering at Seoul National University, Korea, where he is expected to graduate in February 2028. His research interests include renewable energy systems, AI-based seismic and subsurface structure analysis.

Won Joo Lee received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Hanyang University, Korea, in 1989, 1991 and 2004, respectively. Dr. Lee joined the faculty of the Department of Computer Science and Engineering at Inha Technical College, Incheon, Korea, in 2008, where he has served as the Director of the Department of Computer Science and Engineering. He is currently a Professor in the Department of Computer Science and Engineering, Inha Technical College. He has also served as the president of The Korean Society of Computer Information. He is interested in parallel computing, internet and mobile computing, and cloud computing, data science, artificial intelligence, LLM, AI Agents.