오디오 소스 분리 기술을 통한 악보 추적 성능 개선 방안

송상훈*, 한준서*, 김병찬*, 이재혁*, 최선오*, 김순태**

Improving Score Following Performance via Audio Source Separation

Sanghun Song*, Junseo Han*, Byeongchan Kim*, Jaehyuk Lee*, Sunoh Choi*,
and Suntae Kim**

요 약

악보 추적은 연주 중 오디오 신호를 기반으로 악보 내 현재 위치를 실시간으로 추정하는 기술로, 음악 교육과 실시간 반주 등 다양한 분야에 활용되고 있다. 그러나 기존의 실시간 악보 추적 기술은 대부분 합성 음원 데이터를 기반으로 학습되어 실제 연주 환경에서는 다른 악기의 소리나 노이즈, 공간에 따른 소리 변형 등에 의해 성능이 저하되는 한계가 있다. 본 연구에서는 실시간 오디오 분리 기술인 Conv-TasNet을 적용하여 연주 중 발생하는 배경음 및 노이즈를 제거함으로써 실시간 악보 추적 성능을 향상시켰다.

Abstract

Score following is a technology that estimates the current position in a musical score in real time based on the incoming audio signal during a performance. It is applied in various fields such as music education and real-time accompaniment. However, most existing real-time score following methods are trained on synthetic audio data, which leads to performance degradation in real-world environments due to the presence of other instruments, noise, and sound variations caused by acoustic conditions. In this study, we enhance real-time score following performance by applying Conv-TasNet, a real-time audio source separation model, to remove background noise and interference during live performances.

Key words
Score Following, Real performance audio, Multi-Resolution Prediction Model, Conv-TasNet

Ⅰ. 서 론

악보 추적(Score Following)은 연주의 오디오 신호를 기반으로 악보와 컴퓨터 간 실시간 오디오 정렬을 수행하여 현재 연주 위치를 동기화하는 기술이다 [1]. 이 기술은 가상 반주, 자동 페이지 넘김, 음악 교육 및 평가 등 다양한 실시간 음악 응용 서비스에 활용된다. 특히 라이브 연주 환경에서는 연주 흐름을 방해하지 않고 정확한 악보 위치를 인식하는 것이 연주자의 몰입도와 표현력 유지에 매우 중

* 전북대학교, {ssh9199_, hjs123, qudcks3044, messi3000, suno7}@jbnu.ac.kr
** 전북대학교, stkim@jbnu.ack.kr

요하다.

기존에는 블루투스 페이지 터너 페달 장치가 널리 활용되어 왔으나, 원활한 사용을 위해 일정 수준의 숙련 및 적응 과정이 요구되며, 별도의 장비가 추가됨으로써 물리적 부담이 증가한다는 한계를 지닌다. 이에 따라 연주자의 추가적인 동작 없이도 악보 위치를 자동으로 파악할 수 있는 지능형 기술, 즉 악보 추적에 대한 연구가 활발히 이루어지고 있다.

최근에는 악보 이미지와 오디오를 직접 정렬하는 end-to-end 악보 추적 기술이 주목받고 있다. 대표적으로 Henkel et al.과 CPJKU 연구팀의 딥러닝 기반 Multi-Resolution Prediction 모델은 경량 구조와 실시간 처리 성능을 기반으로 다양한 응용 환경에 적합한 성능을 보인다 [2]. 그러나 주로 MIDI 기반의 합성 오디오 데이터를 중심으로 학습되었기 때문에 실제 연주에서는 노이즈나 다른 악기의 소리와 겹칠 경우 정확도가 저하되는 한계가 있다.

이에 본 연구에서는 실제 연주 환경에서의 악보 추적 정확도를 향상시키기 위해 실시간 오디오 소스 분리 기술인 Conv-TasNet [4]을 활용하는 방안을 제안한다. 연주 중 발생하는 음향 간섭, 노이즈를 제거함으로써 모델 입력으로 사용하는 오디오의 품질을 개선하고 이를 통해 기존 모델이 실제 연주 환경에서도 보다 안정적이고 신뢰성 있게 악보를 추적할 수 있도록 한다. 실험 결과, 소리 분리 적용 시 Bar Accuracy는 0.535에서 0.575, System Accuracy는 0.708에서 0.731로 향상되었으며, 프레임 추적 정확도(≤ 1.0초)는 59.5%에서 62.9%로 개선되었다.

Ⅱ. 배경

2.1 Multi-Resolution Prediction 모델 개요

Multi-Resolution Prediction 모델은 실시간 악보 추적을 위해 설계된 경량 딥러닝 기반 모델로, 악보 이미지와 실제 연주 오디오를 정렬하는 작업에 최적화되어 있다. 모델은 크게 두 가지 주요 구성 요소로 이루어진다 [2].

첫째, 악보 이미지 처리 네트워크는 객체 탐지 모델인 YOLO [2]와 유사한 구조를 기반으로 하며, 악보 이미지 내의 음표 객체를 감지하고 각 음표의 좌표 정보를 추출한다. 이 과정에서 CNN 기반 피처 추출기와 바운딩 박스 예측 모듈이 사용되어, 시각적 정보로부터 의미 있는 악보 구조 정보를 효과적으로 추출한다.

둘째, 조건부 순환 신경망은 LSTM을 기반으로 하며, 시간에 따라 변화하는 오디오 신호와 앞서 추출된 음표 좌표 정보를 함께 입력으로 받아 현재 연주 위치를 실시간으로 추정한다. 이 구조는 과거의 연주 흐름과 현재의 오디오 피처를 고려하여, 악보 상에서 어느 부분이 연주되고 있는지를 순차적으로 예측한다.

본 연구에서는 해당 Multi-Resolution Prediction 모델을 기반으로 실험을 진행하며 성능 개선을 시도하였다.

2.2 오디오 소스 분리 기술: Conv-TasNet

오디오 소스 분리는 하나의 오디오 신호에서 서로 다른 소리 성분을 분리해내는 기술로, 음악 정보 처리 및 음성 인식 분야에서 활발히 연구되고 있다. 초기에는 NMF나 ICA, Wiener Filter와 같은 통계적 기법이 사용되었으나 [5], 최근에는 딥러닝 기반의 시간 도메인 분리 모델들이 높은 성능을 보여 주목받고 있다.

특히 Conv-TasNet [4]은 시간 도메인에서 직접 신호 를 처리하는 방식으로, 복잡한 주파수 변환 없이도 빠르고 정확한 소스 분리가 가능하다. 이 모델은 1D 컨볼루션 기반의 인코더-디코더 구조와 마스크 추정 네트워크로 구성되며, 입력 오디오를 잠재 공간으로 인코딩한 후, 소스별 마스크를 적용해 각 소리 성분을 효과적으로 분리하고 디코더를 통해 다시 시간 도메인 신호로 복원한다.

그림 1은 이러한 구조를 시각적으로 보여준다. 혼합 오디오가 인코더를 통해 특징 공간으로 변환된 뒤, 분리 모듈(Separation)에서 각 소스의 마스크가 추정되고, 이후 디코더를 거쳐 각각의 분리된 파형(waveform)이 재구성되는 과정을 포함한다. Conv-TasNet의 핵심 특징은 다음과 같다:

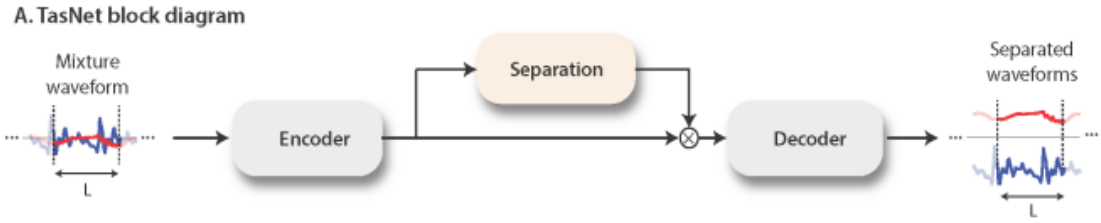

그림 1. Conv-TasNet 블록 다이어그램
Fig. 1. Conv-TasNet Block Diagram

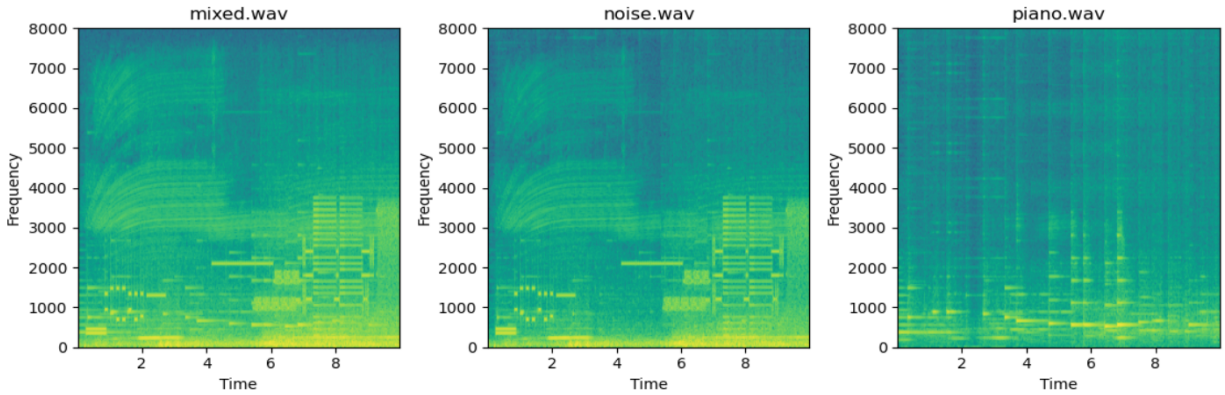

그림 2. Conv-TasNet을 이용한 소리 분리 수행 결과
Fig. 2. Sound separation result of Conv-TasNet

- STFT 변환 없이 직접 시간 도메인에서 동작, 위상 왜곡 없이 소리를 재구성할 수 있다.
- 짧은 프레임 지연과 효율적인 연산량, 실시간 시스템에 적합하다.
- 다양한 환경에서의 잡음 억제 및 악기 분리에 강인한 성능을 보인다.

본 연구에서는 악보 추적 정확도를 높이기 위해 Conv-TasNet을 전처리 단계에 도입하여, 혼합된 연주 오디오에서 피아노 트랙만을 추출하고 이를 모델 입력으로 사용하였다. 이를 통해 실제 연주 환경에서도 보다 안정적이고 신뢰성 있게 악보를 추적할 수 있도록 하였다.

III. 연구 방법

Conv-TasNet [4]를 통해 노이즈가 제거된 피아노 소리를 추출하고, 이를 악보 추적 모델의 입력으로 사용함으로써 최종적으로 연주의 위치를 추적한다.

3.1 오디오 전처리: Conv-TasNet 기반 소스 분리

혼합된 연주 오디오에서 피아노 소리만을 분리 추출하기 위해 Conv-TasNet 모델을 학습하였다. 학습 데이터는 피아노 연주 데이터셋 MAESTRO [7]와 노이즈 데이터셋인 MUSAN [8]을 활용하여 다양한 실환경 조건을 반영하도록 구성하였다. 모든 오디오 샘플을 10초 단위로 분할하고 이를 랜덤으로 혼합한 음원을 생성하였다. 최종적으로 피아노 음원, 노이즈 음원, 혼합 음원, 총 세 종류의 음원이 모델의 입력 샘플로 사용되었으며, 혼합된 입력과 대응되는 피아노 음원을 GT로 구성하였다.

이후 Conv-TasNet은 엔코더-마스킹-디코더 구조를 통해 시간 도메인 상에서 직접 신호를 처리하며, 입력 오디오에서 피아노 성분을 추출하는 방식으로 학습되었다. 학습이 완료된 모델은 실시간 입력 오디오에 대해 피아노 트랙만을 분리하였다.

그림 2는 이러한 과정을 통해 분리해 낸 음원의 스펙트로그램을 나열한 것이다. 분리된 신호를 살펴보면, 혼합된 음원(mixed.wav)을 악보 추적 모델에 그대로 입력할 경우 악보와의 정렬 과정에서 필요한 음악적 이벤트가 다른 소리에 의해 가려져 정확

도 저하 가능성을 짐작할 수 있다. 반면 분리된 피아노 음원(piano.wav)를 사용할 경우 실제 악보와 정렬하기 위한 특징 추출에 유리해진다.

이와 같은 소스 분리 기반 전처리 과정을 통해 실제 연주 환경에서도 다양한 잡음의 영향을 줄이고, 보다 안정적이고 정밀한 악보 추적이 가능하도록 설계하였다.

Ⅳ. 실험 및 결과

4.1 실험 설정

Multi-Resolution Prediction의 사전 학습 모델 중 가장 성능이 뛰어난 것으로 보고된 CYOLO-SB+A을 사용하였다 [2]. 소리 분리가 악보 추적 성능에 미치는 영향을 알아보기 위해 임의의 노이즈를 추가한 MSMD 데이터셋을 기반으로 소리 분리를 적용한 경우와 그렇지 않은 경우로 나누어 CYOLO-SB+A 모델의 입력으로 사용하여 성능 평가를 수행하였다.

성능 평가는 Bar Accuracy, System Accuracy, 그리고 Frame Tracking Ratio를 기준으로 수행하였다.

Bar Accuracy는 마디 단위, System Accuracy는 시스템 단위의 정확도를 의미하며, Frame Tracking Ratio는 시간 오차 범위 내에서 정답과 예측 프레임이 얼마나 일치하는지를 비율로 나타낸다.

4.2 실험 결과

표 1은 소리 분리 전처리의 적용 여부에 따른 악보 추적 성능 차이를 정량적으로 비교한 결과이다.

소리 분리를 적용하지 않은 경우, Bar 단위 정확도는 0.535, System 단위 정확도는 0.708이었다. 프레임 추적 비율(Frame Tracking Ratio)은 오차 기준 0.05초 이하에서 43.4%, 1.0초 이하에서 59.5%, 5.0초 이하에서 77.7%를 기록하였다.

반면, Conv-TasNet 기반의 소리 분리 후에는 Bar 정확도가 0.575로 4.0% 상승하였고, System 정확도는 0.731로 향상되었다. 프레임 추적 비율도 개선되어, 0.05초 이하에서 47.2%, 1.0초 이하에서 62.9%, 5.0초 이하에서는 80.1%로 증가하였다.

이러한 결과는 노이즈나 다른 악기의 영향을 받은 혼합 오디오에서 피아노 신호를 분리함으로써, 악보 추적 모델이 보다 명확한 입력을 받아 성능이 개선되었음을 의미한다. 특히 0.05초에서 1.0초 사이의 오차 범위에서의 추적 정확도 향상은 실시간 연주 상황에서도 보다 안정적인 추적이 가능함을 보여준다.

Ⅴ. 결론 및 향후 연구 방안

본 연구에서는 실제 연주 환경에서의 악보 추적 정확도 저하 문제를 해결하고자, Multi-Resolution Prediction Score Following 모델에 대해 오디오 분리 기술인 Conv-TasNet을 통해 입력 오디오에서 주요 악기 트랙만을 분리하여 노이즈를 제거함으로써 실시간 입력의 품질을 높였다. 이는 본 연구에서 기반으로 한 Multi-Resolution Prediction 뿐 아니라 그 외 자동 페이지 넘김, 연주 분석, 반주 보조 시스템 등의 실제 연주 상황에서 실질적으로 적용할 수 있는 접근으로써 의미가 있다.

실험 결과, 제안한 방식은 노이즈가 있는 환경에서 기존 모델 대비 프레임 추적 정확도 및 onset [2] 기반 추적 정확도에서 모두 성능이 소폭 개선되었다. 본 실험을 진행하는 동안 시간과 GPU 리소스 제약으로 인해 간소한 데이터만을 사용한 점, 실연 데이터인 MAESTRO를 통해서만 학습된 피아노 음원 분리용 Conv-TasNet 모델이 MIDI 합성 음원 데이터로만 이루어진 MSMD를 분리할 때 매우 낮은 분리 성능을 나타낸 점 등의 한계를 고려하면, 훈련 데이터의 규모와 다양성을 확보하여 더욱 큰 성능 향상을 보여줄 것으로 기대된다.

표 1. 실험 결과
Table 1. Experimental Results

지표 항목	프레임 추적 비율 (%)					Bar 정확도 (평균)	System 정확도 (평균)
	≤ 0.05	≤ 0.1	≤ 0.5	≤ 1.0	≤ 5.0		
noise-MSMD	43.4	45.2	54.8	59.5	77.7	0.535	0.708
Sound-separated noise-MSMD	47.2	48.9	58.5	62.9	80.1	0.575	0.731

참 고 문 헌

[1] N. Orio, S. Lemouton, D. Schwarz and N. Schnell, “Score Following: State of the Art and New Developments,” NIME, May 2003

[2] F. Henkel and G. Widmer, “Multi-modal Conditional Bounding Box Regression for Music Score Following,”, May 2021.

[3] A. Galán-Cuenca, J. J. Valero-Mas, J. C. Martinez-Sevilla, A. Hidalgo-Centeno, A. Pertusa, and J. Calvo-Zaragoza, “MUSCAT: a Multimodal mUSic Collection for Automatic Transcription of Real Recordings and Image Scores,” MM 2024, Oct. 2024,

[4] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation,”, May 2019.

[5] H. Sawada, N. Ono, H. Kameoka, and D. Kitamura, “A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF,” APSIPA Trans. Signal Inf. Process., vol. 8, pp. e12, 2019.

[6] M. Dorfer, A. Arzt, and G. Widmer, “Towards Score Following in Sheet Music Images,” Trans. Int. Soc. Music Inf. Retr., vol. 1, pp. 1–13, 2018.

[7] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,”, Oct. 2015

[8] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, “Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset,”, Jan. 2019