Smart Media Journal / Vol.15, No.6 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2026.15.6.36

딥 뉴럴 아키텍처 기반 광학 악보 인식에서 비표준 및 손글씨 악보의 구조적 이해를 위한 최신 방법론에 관한 종합적 고찰

(A Comprehensive Review of Deep Neural Architectures for Optical Music Recognition with Emphasis on Structural Understanding of Non-Standard and Handwritten Music Scores)

정영진*, 나인섭**

(Young Jin Jung , In Seop NA)

요 약

본 논문은 딥 뉴럴 아키텍처 기반 광학 악보 인식(Optical Music Recognition, OMR)의 최신 방법론을 체계적으로 고찰하고, 비표준 및 손글씨 악보의 구조적 이해 문제를 중심으로 기술적 한계와 향후 발전 방향을 분석한다. 이를 위해, 손글씨 악보와 정간보와 같은 비표준 악보에서 요구되는 구조적 추론 능력이 핵심 기술 과제임을 규명하였다. 또한 YOLO 및 Faster R-CNN 기반 객체 검출 접근, CNN-RNN 기반 시퀀스 모델, Transformer 기반 종단간 모델을 구조적 이해 관점에서 비교·분석하고, 최신 Transformer 및 Foundation Model 기반 접근의 특성과 확장 가능성을 함께 고찰하였다. 더불어 Symbol Error Rate(SER), Note-level F1, Sequence Accuracy, MusicXML 유효성 등 주요 평가지표를 통합 정리하고, 각 지표가 반영하는 시각적·음악적·구조적 성능 차이를 분석하였다.

■ 중심어 : 광학 악보 인식 ; 구조적 의미 재구성 ; 비표준 악보 ; 격자 기반 악보 ; 다성 음악 구조

Abstract

This paper presents a systematic review of recent deep neural architecture-based approaches for Optical Music Recognition (OMR), with a particular focus on the structural understanding of non-standard and handwritten musical scores. We identify structural reasoning as a critical challenge in the recognition of non-standard notations, such as handwritten scores and Jeongganbo. To address this, we comparatively examine object detection-based methods (e.g., YOLO and Faster R-CNN), CNN-RNN sequence models, and Transformer-based end-to-end architectures from the perspective of structural understanding. We also discuss recent advances in Transformer and foundation model-based approaches, highlighting their characteristics and potential for generalization. In addition, key evaluation metrics, including Symbol Error Rate (SER), Note-level F1 score, Sequence Accuracy, and MusicXML validity, are systematically reviewed, and their respective roles in assessing visual, musical, and structural performance are analyzed.

■ keywords : Optical Music Recognition (OMR) ; Structural semantic reconstruction ; Non-standard notation systems ; grid-based notation systems ; Polyphonic music layouts

Ⅰ. 서 론

음악은 인류가 보존해야 할 가장 소중한 문화유산 중 하나이다. 이를 기계가 이해할 수 있는 데이터로 변환하는 광학 악보 인식(Optical Music Recognition, OMR) 기술은 디지털 음악학 연구의 근간을 이룬다 [1].

* 정회원, 전남대학교 의공학부
** 종신회원, 전남대학교 문화콘텐츠학부
이 논문은 전남대학교 학술연구비(과제번호: 2025-1041-01) 지원에 의하여 연구되었음.
접수일자 : 2026년 03월 18일
수정일자 : 2026년 04월 10일
게재확정일 : 2026년 04월 14일
교신저자 : 나인섭 e-mail : ypencil@hanmail.net

OMR의 최종 목표는 단순한 이미지 캡처를 넘어 악보 내에 정밀하게 배치된 기호들의 시각적 관계를 해석하고, 이를 통해 음악적 의미(Musical Semantics)를 복원하는 데 있다. 이는 일반적인 광학 문자 인식(Optical Character Recognition, OCR)이 선형적인 문자열(Character Sequence)을 복원하는 것과 달리, OMR은 수직적인 화성 구조와 수평적인 리듬 관계를 동시에 해결해야 하는 고차원적인 2차원 구조화 문제이기 때문이다 [2].

역사적으로 OMR 시스템은 1990년대 초 Musitek의 MIDISCAN과 같은 상용 소프트웨어의 출시와 함께 본격적인 발전을 시작했다 [2]. 초기 시스템은 주로 오선 제거(Staff Removal), 기호 분할 및 분류라는 명시적인 단계적 처리구조(stage-based pipeline)에 의존하는 규칙 기반 알고리즘을 사용하였다. 그러나 이러한 방식은 오선 제거 과정에서 필연적으로 발생하는 정보 유실과, 복잡하게 중첩된 기호들의 분리 실패라는 고질적인 문제에 직면해 왔다. 2010년대 중반 이후 합성곱 신경망과 순환 신경망이 결합된 딥러닝 모델의 등장은 이러한 물리적 분할의 한계를 극복하는 계기가 되었으며, 최근의 트랜스포머 아키텍처는 악보 전체를 하나의 거대한 시퀀스로 이해하는 종단간(End-to-End) 패러다임을 확립하였다 [3]. 현재 OMR 연구는 인쇄된 표준 악보뿐만 아니라, 극심한 변형을 동반하는 손글씨 악보와 서양식 오선보와는 완전히 다른 정간보 등 비표준 체계로 확장되고 있다. 특히, 정간보는 시간 및 공간 표현 방식을 취하는 형태를 가지고 있어 인식의 새로운 접근이 필요하다.

본 논문은 최신 딥 뉴럴 아키텍처가 악보의 구조적 관계를 어떻게 학습하는지 상세히 분석하고, 2025년까지의 최신 벤치마크를 통해 기술적 도달점을 점검한다.

II. 관련 연구

OMR 아키텍처는 과거 규칙 기반 이미지 처리에서 딥러닝으로 넘어오며 객체 검출 기반 접근, CNN-RNN 기반 시퀀스 모델, 그리고 트랜스포머 기반 OMR로 발전하고 있다.

1. 객체 검출 기반 접근

초기 딥러닝 OMR 연구는 악보 내의 각 기호를 독립적인 객체로 간주하는 객체 검출(Object Detection) 기술에 집중했다. YOLO(You Only Look Once)와 Faster R-CNN 모델은 악보 내 수많은 미세 기호들을 바운딩 박스로 빠르게 찾아내는 데 기여했으며, 특히 최신 YOLOv11 모델은 단일 보표뿐만 아니라 전체 페이지 악보에서도 음표와 조표 등을 실시간으로 식별해 낸다 [7]. 그러나 이 접근법은 기호의 위치는 파악하지만 그들 사이의 '음악적 관계성'을 추론하지 못한다. 검출된 기호들을 구조적 포맷으로 묶기 위해서는 복잡한 악보 조립(Notation Assembly) 알고리즘이 필수적이며, 이 과정에서 발생하는 휴리스틱 조립 오류는 전체 시스템의 신뢰도를 급격히 저하시킨다.

2. CNN - RNN 기반 시퀀스 모델

바운딩 박스를 조립하는 후처리의 어려움을 없애기 위해 등장한 CNN-RNN 결합 모델은 악보 이미지를 세로로 얇게 잘라 연속적인 특징을 추출하고, 이를 LSTM/GRU에 입력하여 CTC(Connectionist Temporal Classification) 손실 함수를 통해 시퀀스를 직접 생성한다 [8]. 이 구조는 전처리 과정을 생략한 종단(End to End) 단성 악보 인식에서 우수한 성능을 입증했다. 그러나 RNN 기반 모델은 시퀀스가 길어질수록 첫머리에 등장한 조표를 잊어버리는 등 장거리 의존성(Long-range Dependency)의 한계를 보였으며, 여러 성부가 수직으로 진행되는 다성 음악 구조를 표현하는 데는 구조적 취약성을 노출했다.

3. 트랜스포머 기반 OMR

트랜스포머 아키텍처는 전역 어텐션(Global Attention) 메커니즘을 통해 이미지 내 모든 지점 간의 관계를 동시에 계산함으로써 OMR의 패러다임을 바꿨다 [3]. 멀리 떨어진 마디 사이의 연결성이나 다성 음악의 복잡한 수직적 화음 관계를 모델링하는 데 탁월하다.

2024~2025년에 걸쳐 발표된 핵심 모델 중 “Sheet Music Transformer(SMT)와 SMT++”는 악보 페

이지 전체를 이미지로 입력받아 중간 분할 없이 직접 kern이나 MusicXML 포맷으로 변환하며, 대규모 합성 데이터를 이용한 커리큘럼 학습으로 성능을 극대화했다. 일례로 SMT++ 모델은 복잡한 피아노 폼 데이터셋(GrandStaff 등) 평가에서 기존 파이프라인 대비 에러율을 최대 92~94% 감소시켰으며, 98% 이상의 렌더링 성공률(Renderability)을 기록하며 종단간 인식의 뛰어난 실효성을 입증하였다 [9]. 또한, Llama 3.2 비전 인코더와 같은 대규모 사전학습 모델을 활용한 “Legato(2025)”는 간결한 ABC 표기법을 생성 타깃으로 삼아 극소량의 데이터로도 고도의 일반화와 연산 효율성을 증명하였다 [10].

TrOMR과 Zeus 등의 비전 트랜스포머(Vision Transformer, ViT) 기반 모델들은 복잡한 XML 트리를 선형화한 선형화된 MusicXML(Linearized MusicXML, LMX) 방식을 적용하여 피아노 폼 인식에서 주도권을 확보하고 있다. 특히 TrOMR 모델은 대규모 합성 데이터(MSD)와 실제 카메라 촬영 악보(CMSD) 환경에서 각각 2.5%와 2.4%라는 매우 낮은 기호 오류율(SER)을 달성하여 다성 음악 인식에서 기술적 진보를 보여주었다 [11].

III. 분야적 특성

인공지능(Artificial Intelligence, AI)이 악보를 인식하는 과정은 그 고유의 기호 체계와 암묵적인 음악적 규칙으로 인해 고도의 시각-논리적 매핑 능력을 요구한다. 악보는 서양 오선보의 계층적 구조, 손글씨 악보의 비정형성, 비표준 및 격자 기반 악보 체계 등 다양한 악보 표기 체계와 구조적 특성을 보유하고 있다.

1. 서양 오선보의 계층적 구조

서양 표준 오선보(Common Western Modern Notation)는 그림 1과 같이, 기하학적 규칙성이 매우 뚜렷한 표기 체계로, 5개의 수평 평행선으로 구성된 오선(Staff)이 음높이(Pitch)를 결정하는 수직적 좌표계 역할을 수행한다 [1]. 기호들은 이 좌표계 위에서 엄격한 계층 구조를 형성한다. 첫째, 수평적 흐름에 따라 악보는 왼쪽에서 오른쪽으로 진행되며 마디(Measure) 단위로 시간적 구획이 이루어진다. 둘째, 동일한 수직축 상에 놓인 음표들은 화성(Harmony)을 형성하여 동시에 연주되어야 하는 수직적 동기화를 이룬다. 셋째, 논리적 연결성 측면에서 붙임줄(Tie), 이음줄(Slur), 그리고 빔(Beam)과 같은 기호들은 개별 음표들을 묶어 아티큘레이션(연주 표현)이나 리듬 단위의 새로운 의미를 창출한다. 이러한 복합성 때문에 시각적 기호와 음악적 의미 사이에는 큰 '시맨틱 갭(Semantic Gap)'이 존재한다.

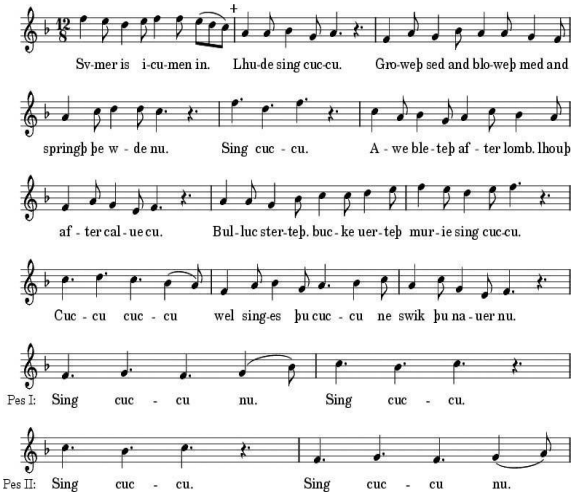

그림 1. 서양 표준 오선보의 예

2. 손글씨 악보의 비정형성

컴퓨터 조판 프로그램으로 작성된 인쇄보와 달리, 손글씨 악보 인식(Handwritten Music Recognition, HMR)은 그림 2와 같이, 심각한 비정형성과 노이즈를 처리해야 하므로 시각적 문맥 분리가 필수적이다. 필기자의 습관에 따라 음표 머리의 크기나 기둥의 기울기가 변하며, 오선 밖으로 삐져나온 덧줄(Ledger lines)이나 조표들이 심하게 중첩되어 분할 자체를 불가능하게 만든다.

특히 최근 주목받는 재즈 리드 시트(Jazz Lead Sheets)의 경우, 멜로디 라인 위에 손글씨로 적힌 화음 기호와 가사가 혼재되어 있어 시각적 문맥 분리가 필수적이다 [4]. 이를 극복하기 위해 최신 연구는 인쇄 악보로 기초 문법을 학습한 후 인위적으로 생성된 변형 노이즈 데이터를 투입하는 '합성 데이터 생성기 기반 커리큘럼 학습(Curriculum Learning)'을 도입하고 있다. 또한, 전문가가 오류 가능성이 가장

높은 지점만 선택적으로 수정하게 하는 능동 학습(Active Learning) 프레임워크를 통해 주석 효율성을 극대화하는 시도가 이어지고 있다 [5].

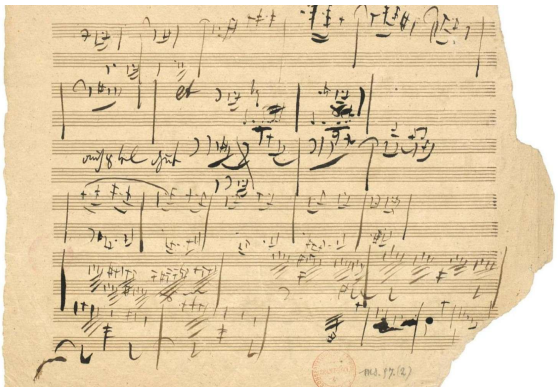

그림 2. 손글씨 악보의 비정형성의 예

3. 비표준 및 격자 기반 악보 체계

서양 오선보에 맞춰 설계된 구조를 전혀 적용할 수 없는 비표준 악보들은 최근 AI 모델의 범용성을 시험하는 주요 무대이다. 15세기 세종대왕 시기에 창제된 한국의 전통 악보인 정간보(Jeongganbo)는 그림 3과 같이, 서양의 오선 대신 우물 정(井) 자 모양의 2차원 상자형 격자를 시간의 기본 단위로 사용한다 [6].

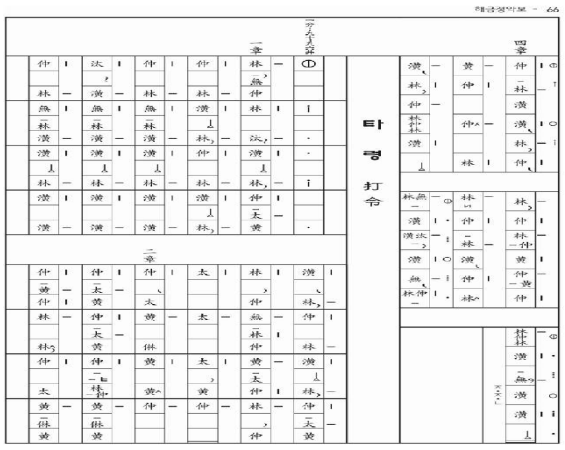

그림 3. 비표준 격자 기반 악보 체계 예[16]

정간보는 위에서 아래로, 오른쪽에서 왼쪽으로 읽는 가독 순서를 가지며, 하나의 정간 내에서 한자(율명)와 부호가 어디에 배치되느냐에 따라 박자의 세부 분할이 결정된다. 이는 1차원 시퀀스로 처리하기 어려운 2차원 격자 해석(Grid Parsing) 문제를 야기한다. 최근 진행된 'Six Dragons Fly Again' 프로젝트는 이러한 특수성을 반영하여 정간 내부의 박자 분할 정보를 수치화 및 토큰화하는 'JG-like encoding' 방식을 제안함으로써 트랜스포머 기반의 성공적인 인식 결과를 도출하였다 [6]. 이 외에도 숫자 표기와 밑줄 리듬 표기가 결합된 중국의 숫자보(Jianpu) 등도 텍스트와 그래픽이 결합된 하이브리드 접근법을 요구한다.

Ⅳ. 구조적 이해와 의미 재구성

OMR의 본질은 기호 검출을 넘어 악보의 '논리적 위계'를 이해하는 것이다. 표 1에서 보는 바와 같이, OMR을 위한 딥 뉴럴 네트워크 기반 모델들은 객체 검출, CNN-RNN 시퀀스, ViT 기반 트랜스포머, Foundation 모델 기반, 비전 트랜스포머 모델 등으로 발전하고 있다. 이들은 표 1과 같은 장점과 한계점을 갖는다. 이들은 악보의 구조적 이해와 의미를 재구성하고 있으며, 특히 딥 뉴럴 아키텍처의 경우 시각적 정보가 단순 문자열이 아닌 논리적 음악 지식 그래프(Music Knowledge Graph)로 변환되어야 완벽한 전사가 이루어진다.

최신 트랜스포머 모델들은 크로스 어텐션 가중치를 통해 출력되는 텍스트 토큰과 이미지 간의 암묵적 정렬을 수행한다. 특히 2025년에 제안된 "LA-SMT(Layout-Aware Sheet Music Transformer)"는 명시적인 레이아웃 주석 데이터 없이도 어텐션 맵 스스로 보표의 경계와 시스템의 구성을 파악하는 암시적 레이아웃 학습(Implicit Layout Learning)을 실현했다 [12]. 이는 모델이 시각 특징을 넘어 음악적 의미가 결합된 내부 표현을 형성하고 있음을 시사한다.

더불어 생성 과정에서의 환각(Hallucination) 오류를 줄이기 위해 구조적 일관성 손실(Structural Consistency Loss)함수가 적극 도입되고 있다. 모델이 예측한 마디 내 박자의 총합이 지정된 박자표와 논리적으로 일치하는지를 검증하여 손실 함수에 패널티로 반영함으로써, 시각적으로는 그럴듯하지만 음악 이론적으로는 불가능한 결과물의 출력을 원천적으로 방지하고 있다. 서양 음악의 시퀀스 규칙이 통하지 않는 비표준 악보의 경우에는 여전히 명시적인 그래프 신경망(Graph Neural Network,

표 1. 딥 뉴럴 아키텍처 비교

모델 유형	주요 논문	장점	한계	대표성능
객체 검출	V. Dvořák [7], A. Pacha [2]	. 실시간 기호 검출 지원, . 시각적 기호 식별 정확도 매우 높음	. 기호 간 음악적 관계 추론 불가, . 복잡한 후처리 조립(Assembly) 알고리즘 필수	- mAP 80~90% 수준 (MUSCIMA++ 및 DeepScores 기준)
CNN-RNN 시퀀스 모델	J. Calvo-Zaragoza [8]	. 수동 규칙이 필요 없는 종단간(End-to-End) 단성 악보 인식	. 시퀀스가 길어질 시 장거리 의존성 소실, . 다성부 화음 구조 표현의 한계	- Sequence-level error (**12.5%** versus 17.9%) - Symbol-level error (**0.8%** versus 1.0%). (PrIMuS 단성보 데이터셋 기준)
Transformer (ViT 기반)	X. Li [3], A. Ríos-Vila [9]	. 전역 문맥 모델링 탁월, . 분할 없는 전체 페이지 다성부 악보 직접 변환	. 막대한 컴퓨팅 자원 필요, . 대규모 사전학습(합성) 데이터에 의존	- TrOMR: SER 2.5% (MSD), 2.4% (CMSD) - SMT++: 에러율 92~94% 감소, 렌더링 성공률 >98% (GrandStaff)
Foundation Model 기반	Y. Jiang [14], G. Yang [10]	. 극소량의 데이터로도 고도의 일반화 가능, . 다양한 포맷(ABC 등) 생성능력 우수.	. 방대한 파라미터 크기로 인한 높은 연산 비용, . 모바일/엣지 환경 적용에 제약	MuFUN 모델: 오류율을 60% 이상 대폭 삭감 또는 기존 모델 대비 정확도 200% Legato: (SMT++) 대비 TEDn 지표는 68%, OMR-NED 지표는 47.6%의 절대 오차 감소 (카메라 촬영본 데이터 기준)
Vision Transformer	Dosovitskiy [4]	. 이미지 토큰화 기반 학습	. 대규모 학습 필요	Medium기준 WER(11.9), CER(13.67), LER(29.68) (JAZZMUS 데이터셋)

GNN)을 활용하여 기호 간 종속성을 엣지로 묶는 융합 연구도 진행 중이다.

Ⅴ. 데이터셋과 평가

OMR 기술이 단순 분류 정확도를 넘어 구조적, 위계적 일관성을 측정하는 방향으로 진화함에 따라, 이를 학습시키고 평가하기 위한 데이터셋과 평가지표 역시 고도화되고 있다. 본 절에서는 OMR 연구의 기반이 되는 주요 데이터셋과 평가 방법론들을 체계적으로 분류하여 서술한다.

1. 주요 데이터셋(DataSets)

OMR 모델의 학습과 평가를 위해서는 고품질의 주석(Annotation)이 포함된 대규모 데이터셋이 필수적이다. 웹을 통해 공개된 대표적인 최신 데이터셋들은 다음과 같다.

- **MUSCIMA++:** 손글씨 악보 인식(HMR) 연구를 위한 가장 대표적인 벤치마크 데이터셋이다. 악보 내의 모든 기호와 이들 간의 노드-엣지(Node-edge) 관계 그래프를 픽셀 단위로 정밀하게 주석 처리하여, 기호 분할 및 문맥적 관계 추론 연구에 널리 활용된다.
- **OmniOMR:** 체코 카렐 대학교(Charles University) 연구진을 주축으로 구축 중인 대규모 데이터셋 프로젝트이다. 최근 기호(Symbol) 개수 면에서 MUSCIMA++를 능가했으며, 다양한 형태의 악보 이미지를 MusicXML 포맷과 정렬하여 제공함으로써 범용적인 종간간 OMR 모델 학습을 지원한다.
- **Sheet Music Benchmark (SMB):** 2025년 새롭게 공개된 685페이지 분량의 대규모 평가 전용 데이터셋이다. 단성부(Monophony)부터 피아노 폼, 4중주(Quartet)에 이르는 다양한 복합 레이아웃을 포함하며, 서양 표준 기보법을 kern포맷으로 제공하여 최신 OMR 기술의 표준화된 성능 평가를 가능하게 한다 [13].
- **특수 목적 데이터셋:** 최근에는 대규모 다중 페이지 인쇄보 생성을 위한 PDMX-Synth데이터셋[10]이나, 멜로디와 화음 기호가 혼재된 293개의 손글씨 악보를 모은 JAZZMUS (Jazz Lead Sheets)[4] 데이터셋 등 특정 도메인에 특화된 자료들이 활발히 구축되고 있다.

이러한 발전에도 불구하고 정간보나 계량 기보법 등 비표준 악보 데이터셋의 절대적인 부족은 범용 OMR 연구 확장의 가장 큰 제약 요인으로 지적되고 있다.

2. 주요 평가 방법

OMR 성능 평가는 단순 “정확도(accuracy)”로

는 충분하지 않으며, 기호 단위 오류, 음표 단위 의미 정확성, 시퀀스 전체 일관성, 구조적 포맷 적합성을 각각 다르게 측정해야 한다. 아래에서는 OMR 연구에서 널리 사용되는 네 가지 대표 지표를 수식과 함께 체계적으로 설명한다.

가. Symbol Error Rate (SER)

SER은 예측된 기호 시퀀스와 정답 시퀀스 간의 레벤슈타인 거리(Edit Distance)를 기반으로 삽입, 삭제, 대체 오류의 총합을 전체 기호 수로 나눈 오류율이다. SER은 OMR 시스템을 텍스트 번역기와 동일한 관점(구문론적 관점)에서 평가하는 원초적인 지표이다. 이 수치가 낮다는 것은 모델이 이미지 내의 시각적 토큰(점, 선, 기호)들을 놓치지 않고 눈으로 정확히 "읽어냈다"는 것을 의미한다. 모델의 전반적인 비전(Vision) 성능을 대략적으로 가늠하는 훌륭한 기준점이 된다. SER(Symbol Error Rate)은 예측된 기호 시퀀스와 정답 시퀀스 간의 편집 거리(Edit Distance) 기반 오류율이다. 이는 OCR의 CER(Character Error Rate)와 유사한 개념이다.

$$SER = \frac{S + D + I}{N}$$

S: Substitutions (대체 오류)
D: Deletions (삭제 오류)
I: Insertions (삽입 오류)
N: 정답 기호 개수

특징을 살펴보면, 기호 단위 평가를 진행하고, 순서 정보를 반영하며, 리듬·음높이 의미는 고려하지 않는다.

표 2. SER의 장·단점

장점	단점
모델 간 비교 용이	의미적 정확성 반영 부족
시퀀스 오류 반영	구조적 문법 오류 구분 어려움

표 2에서와 같이 SER은 모델간 비교가 쉽고, 시퀀스 오류를 반영한다는 장점이 있지만, 의미적 정확성 반영이 부족하고 SER가 낮아도 음표의 길이나 화음 구조가 틀릴 수 있다. 따라서, SER만으로 이 모델이 '음악을 이해했다'고 단정 지을 수 없다. 예를 들어, 모델이 악보에 묻은 얼룩을 스타카토 점으로 오인해 추가(삽입 오류)하는 것과, 마디 맨 앞에 있는 조표(Key Signature, 플랫이나 샵) 하나를 빼먹는(삭제 오류) 것은 SER 수치상으로는 동일하게 '1개의 에러'로 취급된다. 하지만 후자의 경우, 그 마디에 있는 모든 음표의 음높이가 반음씩 틀려버리는 치명적인 음악적 오류를 갖는다. 따라서 SER은 모델의 기초 체력을 확인하는 지표일 뿐, 최종 결과물의 음악적 가치를 보장하지 못한다.

나. Note-level F1 점수

SER의 형태적 한계를 보완하기 위해 도입된, 음악의 의미(Semantics) 중심 평가 지표이다. 이 지표는 시각적인 기호의 나열이 아니라, 악보가 지시하는 실제 '소리'의 결과물에 초점을 맞춘다. 개별 음표가 가지는 "음높이(Pitch), 시작 시점(Onset), 길이(Duration)"라는 3대 요소가 정답과 완벽히 일치해야만 정답(True Positive)으로 인정한다. Precision(정밀도)과 Recall(재현율)의 조화 평균을 통해 모델이 생성해 낸 음표들의 신뢰도를 산출한다. Note-level F1 점수는 음표 단위(피치 + 시작 시점 + 길이) 일치 여부를 기준으로 계산한다.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2PR}{P + R}$$

TP: 정확히 맞춘 음표
FP: 잘못 예측한 음표
FN: 놓친 음표

표 3. Note-level F1 점수 장·단점

장점	단점
음악적 의미 반영	구조 문법 오류는 반영 어려움
화음 평가 가능	MusicXML 구조 오류 미반영

표 3에서와 같이 음높이와 시간 정보를 반영할 수 있어, 음악적 의미를 반영하고, 화음 평가가 가능한

장점을 갖는다. 다만, 구조 문법 오류는 반영이 어렵고, MusicXML 구조 오류 역시 반영이 안 되는 단점을 갖는다. 따라서, Note-level F1 점수는 실제 연주 재현 정확성과 밀접하다. 여기서, Note-level F1 점수가 높다는 것은 모델이 생성한 악보를 컴퓨터(MIDI)나 연주자가 그대로 연주했을 때 원본과 거의 동일한 음악이 흘러나온다는 것을 의미한다(Replayability 확보). 모델이 음표의 기둥(Stem)을 위로 그릴지 아래로 그릴지 시각적 레이아웃을 헷갈렸더라도, 그것이 내는 소리(음높이와 길이)만 맞다면 이 지표에서는 감점되지 않는다. 따라서 음악 정보 검색(MIR), 자동 반주 트랙 생성, 자동 조옮김(Transposition) 등 실용적인 음악 애플리케이션에 이 모델을 즉시 투입할 수 있는지를 판단하는 가장 실질적이고 핵심적인 척도로 해석할 수 있다.

다. 시퀀스 정확도(Sequence Accuracy)

예측된 전체 기호 시퀀스가 정답 시퀀스와 단 하나의 표기나 오차도 없이 100% 동일한 샘플의 비율을 측정하는 가장 가혹하고 엄격한 평가 기준이다. 예측 시퀀스가 정답 시퀀스와 완전히 동일할 확률을 표현한다.

$$Sequence\ Accuracy = \frac{\text{완전히 일치한 샘플}}{\text{전체 샘플}}$$

표 4. 시퀀스 정확도의 장·단점

장점	단점
완전 일치 여부 명확	매우 낮게 나오는 경향
구조 오류에 민감	부분 정확성 반영 불가

표 4에서처럼, 시퀀스 정확도는 엄격한 평가 방식으로 트랜스포머(Transformer) 기반 모델에서 자주 사용된다. 시퀀스 정확도는 구조 오류에 대해 잘 판별하고, 완전히 샘플이 일치하는 여부를 명확히 알 수 있는 장점이 존재하지만, 복잡한 악보에서는 이 값이 급격히 낮아질 수 있다.

이 지표는 트랜스포머와 같은 종단간 생성 모델의 '안정성'과 '오류 전파(Error Propagation)' 정도를 진단하는 데 핵심적인 역할을 한다. 자기회귀(Auto-regressive) 방식으로 작동하는 모델은 앞에서 한 번 실수를 하면 뒤이어 나오는 예측이 도미노처럼 무너지는 경향이 있다. 99%의 SER 정확도를 보이는 모델이라도 Sequence Accuracy가 현저히 낮다면, 그 모델은 실전에 투입되었을 때 페이지마다 한두 개의 치명적 결함을 꾸준히 발생시킨다는 의미다. 특히 화음이 겹치는 다성부 악보에서는 완벽하게 일치할 확률이 기하급수적으로 떨어지므로, 모델이 복잡한 구조를 얼마나 환각(Hallucination) 없이 견고하게 생성해 내는지를 확인하는 스트레스 테스트 지표로 기능한다.

라. MusicXML 유효성

OMR 모델(특히 트랜스포머 디코더)이 텍스트 형태로 직접 생성해 낸 MusicXML이나 MEI 코드가, XML 표준 문법과 스키마(Schema)를 준수하고 있는지를 검증하는 구조적 지표이다. MusicXML의 유효성(Structural Validity)은 다음 세 가지의 항목을 점검한다. 1) 생성된 MusicXML 파일이 XML 문법적으로 유효한가? 2) MusicXML DTD/Schema를 만족하는가? 3) 음악 소프트웨어에서 정상 파싱 가능한가?

표 5. MusicXML 유효성의 장·단점

장점	한계
실제 사용 가능성 평가	기호 정확도 직접 반영 안함
문법 오류 탐지	의미 오류 미검출 가능

표 5에서와 같이, 유효성 검증을 통해 실제 사용 가능성과 구조적 문법 오류 검출한다. 다만, 기호 정확도가 직접 반영되지 않고, 의미 오류에 대해서 검출되지 않을 가능성이 있는 한계가 있다. 즉, XML이 유효하더라도 음표 길이가 틀릴 수 있다. 인공지능이 악보를 아무리 정확한 소리로 전사(Transcription)했다 하더라도, 괄호 태그(<note>)를 닫지 않았거나 계층 구조를 어겼다면 해당 파일은 껍데기에 불과하다. 유효성이 낮다는 것은 결과물을 MuseScore나 Finale 같은 실제 상용 사보 프

로그램에서 열 수 없으며(Parsing 에러 및 크래시 발생), 인간 편집자가 후수정을 하려고 해도 파일을 로드할 수조차 없음을 의미한다. 이를 통해 우리는 최신 딥러닝 OMR 연구가 단순히 정답을 맞히는 것을 넘어, '인간-기계 상호작용(HCI)'이 가능한 실용적이고 규격화된 소프트웨어 산출물을 만들어낼 능력이 있는지를 명확히 평가할 수 있다.

마. OMR 평가지표 비교 및 연구 적용 전략

앞서 살펴본 바와 같이, 단일 지표만으로는 OMR 시스템의 성능을 온전히 대변할 수 없다. 각 지표는 표 6에서 보는 바와 같이, 시각, 청각, 구조라는 서로 다른 관점에서 모델을 비교 분석할 수 있다.

표 6. OMR 평가지표 비교

지표	평가 수준	순서 반영	의미 반영	구조 문법 반영	엄격도
SER	시각적 기호(토큰) 단위	O	X	X	보통 (기초 시각 인식력 평가)
Note F1	음표(소리의 의미) 단위	부분적	O	X	높음 (실제 연주 가능성 평가)
Sequence Acc.	전체 시퀀스 단위	O	O	부분적	매우 높음 (생성 안정성 및 오류 전파 평가)
MusicXML Validity	구조 및 포맷 단위	X	X	O	구조 중심 (소프트웨어 호환성 평가)

앞서 언급한 지표 해석을 바탕으로, 연구의 목적과 해결하고자 하는 과제의 난이도에 따라 평가 지표를 아래와 같이 입체적으로 조합하는 전략적 접근이 요구된다.

표준 인쇄 악보 전사: SER + Note-level F1조합을 사용한다. 노이즈가 적은 환경이므로, 기호를 얼마나 잘 보았는지(SER)와 그것이 정확한 음악으로 치환되었는지(Note F1)의 균형을 증명하는 데 집중한다.

- **손글씨 악보 (HMR):** SER + Note-level F1 + Sequence Accuracy조합이 필수적이다. 필기체의 극심한 형태 변형과 잉크 번짐 속에서도 모델이 붕괴하지 않고 전체 문맥을 유지해 내는 강건성(Robustness)을 Sequence Accuracy를 통해 추가로 입증해야 한다.
- **종단간(End-to-End) 포맷 변환 (MusicXML 생성):** Sequence Accuracy + MusicXML 유효성조합을 최우선으로 둔다. 사람이 직접 파일을 열람하고 수정할 수 있는 완전한 파일 구조를 생성하는 것이 목표이므로, 문법적 무결성이 곧 시스템의 완성도를 의미하기 때문이다.
- **비표준 악보 (정간보, 숫자보 등):** 서양 오선보 체계와는 완전히 다른 2차원 공간 및 시간 표현 방식을 가지므로, 기존 지표에 더해 각 악보 고유의 문법과 배열 규칙을 평가할 수 있는 "구조적합성 지표(Structural Suitability Metrics)"의 추가 적용이 필수적이다.

바. 최신 연구 분석 (2023 - 2025)

최근 OMR 연구는 '기호를 얼마나 잘 읽었는가(SER)'라는 1차원적 질문을 넘어, "모델이 악보의 위계적 구조와 음악 이론을 얼마나 깊이 있게 이해하고 있는가?"를 묻는 방향으로 진화하고 있다.

- **TEDn과 OMR-NED의 부상:** 기존의 1차원 선형 편집 거리(SER)가 가진 음악적 맹점을 극복하기 위해, 악보를 다차원 트리 구조로 간주하고 노드 간의 유사도를 측정하는 TEDn(Tree Edit Distance with note flattening)[11]이 대세로 자리 잡고 있다. 나아가 음표 머리, 빔, 음높이 등 음악적 비중에 따라 기호의 에러 가중치를 다르게 부여하는 OMR-NED(OMR Normalized Edit Distance)지표가 2025년 SMB(Sheet Music Benchmark) 데이터셋과 함께 새로운 평가 표준으로 채택되었다 [13]. 이는 평가 지표 자체가 인간 음악가의 채점 방식을 모방하기 시작했음을 보여준다.
- **이론 지식의 주입 (Structural Consistency Loss):** 모델 평가를 넘어, 학습 과정 자체를 혁신하는 시도가 이루어지고 있다. 모델이 마디 내의 박자 총합을 틀리거나, 4/4박자에 맞지 않는 엉뚱한 리듬을 생성할 경우 페널티를 부여하는 '구조적 일관성 손실(Consistency Loss)' 함수가 도입되었다. 이는 신경망에 단순히 패턴 매칭을 시키는 것을 넘어 규칙 기반의 음악 이론

(Music Theory)을 강제로 학습시키는 신경-기호(Neuro-symbolic) AI로의 전환을 의미하며, 이를 통해 환각 현상을 획기적으로 줄이고 생성 결과의 신뢰성을 극대화하고 있다.

Ⅵ. 향후 연구 과제

1. 비표준 악보 전용 모델 개발

서양 오선보에 편중된 모델은 한국의 정간보, 중국 숫자보, 고대 성가 악보 등에서 오작동을 일으킨다. 이를 해결하기 위해 단순 1차원 위치 인코딩을 넘어 이미지 내 행렬 정보를 주입하는 2D 위치 인코딩(Grid Parsing) 기술과, 각 비표준 악보 문법에 맞는 음악 특화 커스텀 토크나이저 개발이 요구된다. 최근 정간보 연구에서 제안된 'Beat Counter' 임베딩처럼 시간적 위치를 명시적으로 제어하는 구조를 타 비표준 모델에도 적용한다면, 박물관에 소실 위기로 방치된 수만 장의 고악보를 MIDI(Musical Instrument Digital Interface) 음원으로 완벽히 복원하는 문화유산 복원 솔루션을 구축할 수 있다 [6].

2. 대규모 사전학습을 활용한 OMR 모델

새로운 스타일의 악보가 등장할 때마다 방대한 주석 데이터셋을 구축하는 것은 비효율적이다. 최근 언어 모델의 거대 트렌드를 이식한 MuFun이나 MuCUE와 같은 오디오-비전 기초 모델(Foundation Model) 연구가 가속화되고 있다 [14]. 마스크 오토인코더(Masked Autoencoder, MAE) 등 자기 지도 학습을 활용해 레이블 없이 수백만 장의 악보 이미지와 음악적 시각 특징을 사전 학습하고, 이를 비전-언어 모델(Vision-Language Model, VLM)과 결합한다면, 본 적 없는 창작 기보법이라도 모델 스스로 패턴 유추해 표준 디지털 악보로 번역해 내는 제로샷(Zero-Shot) 악보 번역기의 탄생을 기대할 수 있다 [10].

3. 멀티모달 학습

잉크가 지워졌거나 심하게 훼손된 악보는 시각 정보만으로 정확한 피치를 판별하기 어렵다. 시각적 한계를 돌파하기 위해 악보 이미지(OMR)와 자동 채보(Automatic Music Transcription, AMT) 기술을 결합하는 멀티모달 프레임워크가 부상하고 있다. 대조 학습(Contrastive Learning)이나 동적 시간 워핑(Dynamic Time Warping, DTW) 알고리즘을 통해 시각적 기호 시퀀스와 오디오 파형 신호를 실시간 정렬하면, 모호하게 번진 음표를 소리 피치 정보로 상호 보정하여 100%에 가까운 인식률을 달성할 수 있다. 이는 곧 연주자의 소리를 듣고 실시간으로 스마트 기기의 악보 페이지를 넘겨주거나 틀린 음을 짚어주는 인텔리전트 음악 교육 솔루션으로 직결된다.

4. 모바일 환경 적응

실제 사용자는 통제된 스캔본이 아니라 조명 반사, 페이지 굴곡, 그림자가 존재하는 스마트폰 카메라 사진을 OMR에 입력한다. 이러한 도메인 격차를 해소하기 위한 생성적 데이터 증강(Generative Data Augmentation) 등 강건한 도메인 적응 기법이 요구된다. 또한 실시간 서비스를 위해 Moises-Light와 같은 지식 증류(Knowledge Distillation) 기반의 초경량 딥러닝 아키텍처 개발이 필수적이다 [15]. 이를 통해 인터넷 연결 없이도 스마트폰 카메라로 악보를 비추면 즉석에서 이조(Transposition)를 수행하고 파트별로 분리하여 태블릿 화면에 렌더링해 주는 On-Device 기반 합주용 라이브 악보 스캐너 환경을 구현할 수 있다.

Ⅶ. 결 론

딥 뉴럴 아키텍처의 비약적인 발전은 OMR을 단순한 기호의 나열 및 판독 시스템에서 벗어나, 고도의 구조적 이해를 갖춘 음악 지능(Musical Intelligence) 시스템으로 변모시켰다. 트랜스포머 기반의 종간간 모델은 수직적 화음과 수평적 리듬이 얽힌 다성 악보의 복잡한 계층 구조를 전역 문맥 안에서 성공적으로 포착하며 인식률의 한계를 갱신하고 있다. 특히 2025년 기준의 SMB와 OMR-NED, TEDn과 같은 최신 평가 지표들은 이러한 OMR 기

술의 성숙도를 보다 음악적이고 정밀하게 검증할 수 있는 탄탄한 기반을 마련해주었다.

차세대 OMR 연구는 특정 문화권의 표준화된 악보 체계를 넘어, 정간보나 손글씨 리드 시트와 같이 비정형화된 모든 음악적 유산을 포괄하는 범용 파운데이션 모델로 나아가고 있다. 이러한 기초 모델의 뛰어난 일반화 성능과, 오디오 신호를 융합하여 시각적 불확실성을 극복하는 멀티모달 학습 기술은 전 세계 도서관에 잠들어 있는 방대한 악보를 온전한 디지털 자산으로 전환하는 핵심 엔진이 될 것이다. 궁극적으로 이 기술은 디지털 음악학 연구를 혁신하고, 시각 장애 음악인을 위한 접근성을 개선하며, 인간의 음악 창작 및 교육 패러다임을 한 차원 높은 곳으로 이끌 것이다.

감사의 글: 저자들은 본 논문의 번역 및 언어적 정교화 작업에 있어 Google Gemini(Canvas)의 도움을 받았음을 밝힙니다

REFERENCES

[1] J. Calvo-Zaragoza, J. Hajič jr., and A. Pacha, "Understanding Optical Music Recognition," *ACM Comput. Surv.*, vol. 53, no. 4, pp. 1-35, 2020.
[2] A. Pacha et al., "A Baseline for General Music Object Detection with Deep Learning," *Appl. Sci.*, vol. 8, no. 9, p. 1488, 2018.
[3] X. Li et al., "TrOMR: Transformer-Based Polyphonic Optical Music Recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2023.
[4] C. Wick et al., "Optical Music Recognition of Jazz Lead Sheets," arXiv preprint arXiv:2509.05329, 2025.
[5] J. P. Martinez-Esteso et al., "Human vs. Machine: Comparing Selection Strategies in Active Learning for Optical Music Recognition," Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR), 2025.
[6] D. Jeong et al., "Six Dragons Fly Again: Reviving 15th-Century Korean Court Music with Transformers and Novel Encoding," Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR), 2024.
[7] V. Dvořák, "Fast Optical Music Recognition Using the YOLO Platform," *Digital Repository of Charles University*, 2024.
[8] J. Calvo-Zaragoza, L. Micó, and J. Oncina, "End-to-End Neural Optical Music Recognition of Monophonic Scores," *Appl. Sci.*, vol. 8, no. 4, 2018.
[9] A. Ríos-Vila, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, "End-to-End Full-Page Optical Music Recognition for Pianoform Sheet Music," *arXiv preprint arXiv:2405.12105v4*, 2024 (Updated 2025).
[10] G. Yang et al., "LEGATO: Large-scale End-to-end Generalizable Approach to Typeset OMR," arXiv preprint arXiv:2506.19065, 2025.
[11] E. Cervantes et al., "A Unified Representation Framework for the Evaluation of Optical Music Recognition Systems," arXiv preprint arXiv:2312.12908v2, 2023.
[12] A. Ríos-Vila et al., "An implicit layout-aware transformer for full-page end-to-end optical music recognition," *Int. J. Multimed. Inf. Retr.*, 2025.
[13] J. C. Martinez-Sevilla et al., "Sheet Music Benchmark: Standardized Optical Music Recognition Evaluation," Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR), 2025.
[14] Y. Jiang et al., "Advancing the Foundation Model for Music Understanding," arXiv preprint arXiv:2508.01178v1, 2025.
[15] Moises Research Innovations, "From Separation to Creation: Moises Research Innovations 2025," *Music AI*, 2025.
[16] D. Kim, D. Han, D. Jeong, J.J. Valero-Mas On the automatic recognition of jeongganbo music notation: dataset and approach *J. Comput. Cult. Herit.*, vol. 18, no. 3, pp. 1-21, 2025.

저 자 소 개

정영진(정회원)
2005년 금오공과대학교 전자공학과 학사 졸업
2007년 경북대학교 의용생체공학과 석사 졸업
2011년 연세대학교 의공학과 박사 졸업.
2026년 현재 전남대학교 교수(의공학과)
<주관심분야 : 의료인공지능, 뇌공학, 뉴로모듈레이션>

나인섭(종신회원)
1997년 전남대학교 전산학과 학사 졸업
1999년 전남대학교 전산통계학과 석사 졸업
2008년 전남대학교 전산학과 박사 졸업
2026년 현재 전남대학교 교수(문화콘텐츠학부, 데이터사이언스대학원, 컴퓨터공학대학원)
<주관심분야 : 인공지능, 시각지능, 영상처리, 패턴인식 객체(검출,분할,인식,추적,이해)>