크로스 게이팅 기반 오디오-악보 정보 융합을 활용한 바이올린 운지법 생성

여예린, 이승민, 김정현
세종대학교
yerinee0949@sju.ac.kr, alice20130712@gmail.com, j.kim@sejong.ac.kr

Violin Fingering Generation with Cross Gating-based Audio-Symbolic Fusion

Yerin Yeo, Seungmin Lee, Junghyun Kim
Sejong Univ.

요 약

본 논문에서는 SE Block을 활용하여 심볼릭과 오디오 각각의 중요한 특징을 강조 및 추출하며, Cross Gating으로 두 특징 간 정보를 효과적으로 공유하며 융합하는 향상된 바이올린 운지법 생성 모델을 제안한다. 또한, 데이터의 클래스 불균형 문제를 완화하며 학습 안정성을 높이기 위해 Focal Loss와 ReduceLROnPlateau 학습률 스케줄러를 적용하였다. 실험 결과, 제안 모델은 기존 모델 대비 모든 평가 지표에서 성능이 개선되었으며, 이를 통해 제안 모델의 높은 일반화 능력을 확인하였다.

Ⅰ. 서 론

바이올린 운지법 생성은 주어진 음표 시퀀스를 각 음에 대응하는 현, 위치, 손가락 조합으로 매핑하는 과정으로, 연주자의 취향과 음악적 맥락 등 다양한 요소에 영향을 받는다. 기존 연구 [1 - 2]에서는 심볼릭(symbolic) 데이터를 기반으로 운지법을 예측하는 모델을 제안하였으나, 연주자의 개별적 표현과 실제 연주에서 나타나는 오디오(Audio) 정보를 충분히 반영하지 못하는 한계가 있었다. 이러한 한계를 보완하기 위해 최근 연구 [3]에서는 심볼릭 데이터와 오디오 데이터를 융합하여 학습하는 Aud-Sym 융합 모델을 제안하였다.

기존 Aud-Sym 융합 모델 [3]의 간단한 임베더 구조와 단순한 연접 방식은 각 특징의 중요성과 두 특징 간 상호작용을 충분히 반영하지 못한다는 한계를 가진다. 본 논문에서는 이러한 한계를 보완하기 위해 SE Block [4]과 Cross Gating 구조를 적용하여 운지법 생성 성능을 향상시켰다. 또한, 데이터에 존재하는 클래스 불균형 문제를 해결하기 위해 Focal Loss [5]와 ReduceLROnPlateau 학습률 스케줄러를 적용하였다.

Ⅱ. 본론

본 논문에서 제안하는 바이올린 운지법 생성 모델은 바이올린 연주곡의 오디오 특징과 이로부터 추출된 심볼릭 특징으로 구성된 길이 32의 시퀀스 데이터를 입력으로 사용한다. 오디오와 심볼릭 특징은 각각 SE Block을 포함한 임베딩 모듈에 의해 중요한 정보가 강조되고 추출된다. 이후 Cross Gating에 의해 각 특징에 대한 게이트 값이 계산되며, 해당 값이 상대 특징에 곱해짐으로써 두 특징 간의 상호작용이 효과적으로 반영되고 융합된다. 이렇게 융합된 시퀀스 특징은 곡의 시간적 흐름을 반영하기 위해 bidirectional LSTM(BiLSTM)에 입력되며, BiLSTM의 출력은 두 개의 Linear 층을 거쳐 최종 운지법 클래스 예측에 사용된다. 레이블은 4개의 현, 12개의 위치, 5개의 손가락 조합으로 구성된 총 240개 클래스로 구성된다. 제안 모델의 전체 구조는 그림 1에 나타나 있다.

본 논문에서는 실험을 위해 TNUA와 YTVF 두 가지 바이올린 연주 데이터셋을 사용하였다. TNUA 데이터셋 [1]은 14곡의 악보를 10명의 바이올리니스트가 연주한 140개의 녹음으로 구성되며, 총 113,208개의 음표로

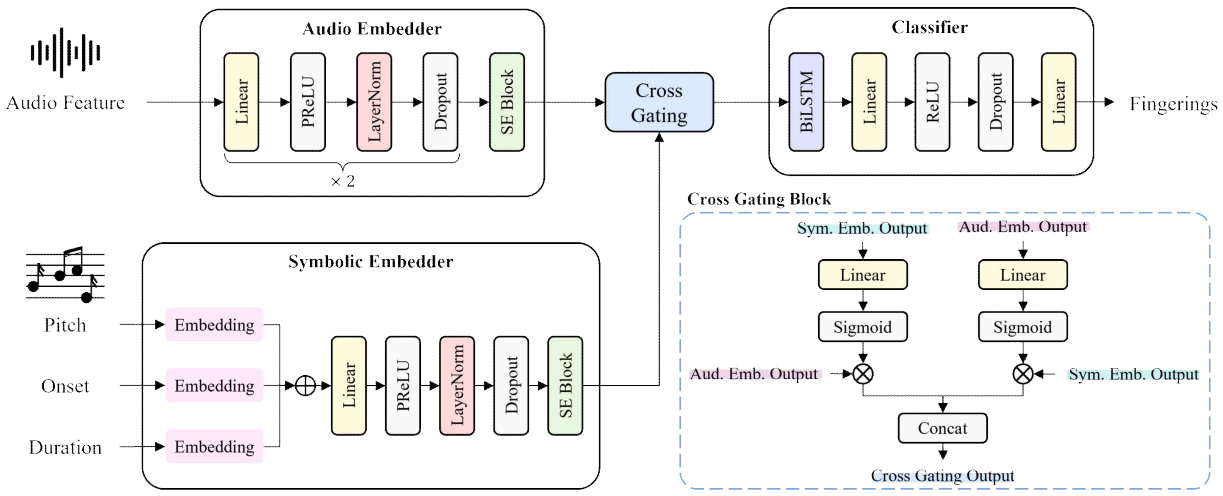

그림 1. 제안하는 바이올린 운지법 생성 모델

구성된다. YTVF 데이터셋 [3]은 공연 영상으로부터 수집된 30개의 연주 영상과 20,562개의 음표로 구성되며, 연주자의 운지법을 수작업으로 작성한 주석을 포함한다. 두 데이터셋 모두 22,050Hz로 샘플링된 오디오와 심볼릭 데이터를 제공하며, 오디오 특징은 소스 분리 및 스펙트로그램 변환을 통해 추출된 프레임별 특성의 평균을 취한 값이다. 심볼릭 특징은 각 음표의 음높이를 나타내는 MIDI 번호인 pitch, 음표의 시작 시점을 나타내는 onset, 음표의 길이를 나타내는 duration 정보로 구성된다.

실제 바이올린 연주에서는 특정 현, 위치, 손가락 조합이 빈번하게 사용되기 때문에, 그림 2와 같이 데이터 내에 클래스 불균형이 발생한다. 이러한 문제를 완화하고자 본 논문에서는 모델 학습 시, Focal Loss를 적용하였다. Focal Loss는 일반적인 Cross-entropy Loss에 비해 잘 분류되는 샘플의 손실 기여도를 줄이고, 소수 클래스의 기여도를 상대적으로 증가시켜 모델이 소수 클래스에 더 집중하도록 만든다. 이와 더불어, 모델의 학습 안정성을 강화하기 위해 ReduceLROnPlateau 학습률 스케줄러를 적용하였으며, Adam 옵티마이저와 초기 학습률 0.001로 학습을 진행하였다. ReduceLROnPlateau는 검증 손실이 일정 epoch 동안 개선되지 않으면 학습률을 동적으로 감소시켜 안정적인 수렴을 유도한다.

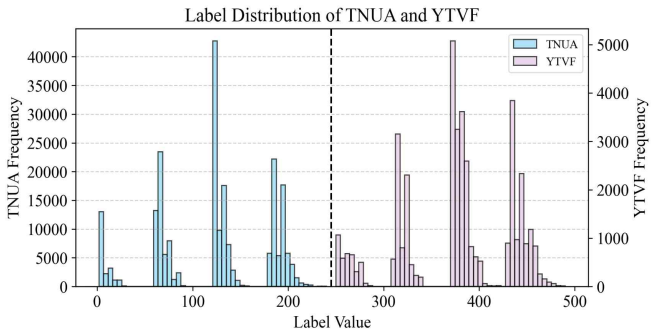

그림 2. TNUA와 YTVF 데이터 셋의 클래스 분포

본 논문에서는 제안 모델의 성능을 평가하기 위해 두 가지 실험, Cross-violinist와 Cross-dataset을 수행하였다. Cross-violinist 실험은 TNUA 데이터셋 내에서 연주자와 곡이 겹치지 않도록 학습 및 평가 데이터를 분리하여 처음 본 연주자와 곡에 대한 일반화 성능을 평가하고, Cross-dataset 실험은 TNUA 데이터셋으로 학습하고 YTVF 데이터셋으로 테스트하여 데이터셋 간의 일반화 성능을 평가하였다. 성능 평가 지표로는 Total accuracy(ACC), Mean Reciprocal Rank(MRR), F1-score를 사용하였다. Total ACC는 예측된 현, 위치, 손가락이 모두 정답과 일치하는 비율이며, MRR은 정답의 예측 순위에 대한 평균 역수, F1-score는 클래스별 Precision과 Recall의 조화 평균을 나타낸다.

실험 결과, 제안 모델은 Cross-violinist 실험과 Cross-dataset 실험 모두에서 기존 모델 대비 모든 평가 지표에서 성능이 향상되었다. 특히 Cross-violinist 실험에서는 Position F1-score가 4.04%, Finger F1-score가 2.64% 개선되었으며, 자세한 결과는 표 1에 나타나 있다.

표 1. Cross-violinist 실험 결과

Model		기존 모델[3]	제안 모델
Total ACC		0.547	0.558
String	MRR	0.901	0.913
	F1-score	0.795	0.820
Position	MRR	0.760	0.771
	F1-score	0.322	0.335
Finger	MRR	0.757	0.765
	F1-score	0.606	0.622

Cross-dataset 실험에서도 제안 모델은 기존 모델 대비 Position F1-score가 15.00%, Finger F1-score가 3.32% 향상되었으며, 자세한 결과는 표 2에 제시되어 있다. 이러한 실험 결과를 통해 제안 모델이 처음 접하는 연주자와 곡, 다른 데이터셋에 대해서도 우수한 일반화 성능과 적응력을 보임을 확인하였다.

표 2. Cross-dataset 실험 결과

Model		기존 모델[3]	제안 모델
Total ACC		0.512	0.526
String	MRR	0.892	0.903
	F1-score	0.768	0.794
Position	MRR	0.722	0.726
	F1-score	0.220	0.253
Finger	MRR	0.742	0.750
	F1-score	0.573	0.592

Ⅲ. 결론

본 논문에서는 SE Block과 Cross Gating 구조를 적용하여 바이올린 운지법 생성 성능을 향상시키고, Focal Loss와 ReduceLROnPlateau 학습률 스케줄러를 도입하여 학습 안정성을 고려한 바이올린 운지법 생성 모델을 제안하였다. 제안 모델은 Cross-violinist 및 Cross-dataset 실험에서 기존 모델 대비 모든 평가 지표에서 성능이 향상되었으며, 처음 접하는 연주자와 곡, 다른 데이터셋에 대해서도 일반화 능력을 확인할 수 있었다. 이러한 결과를 바탕으로, 향후 더 다양한 연주자 데이터를 활용하고 실제 연주 상황에 대한 평가를 통해 모델의 실시간 적용 가능성을 검증하고자 한다. 이를 통한 연주 지원 시스템과 음악 교육 도구 등 다양한 응용 분야로의 확장이 기대된다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(메타버스융합대학원)사업 연구 결과로 수행되었음 (IITP-2023-RS-2023-00254529).

참 고 문 헌

[1] Yi-Hsin Jen, Tsung-Ping Chen, Shih-Wei Sun, and Li Su, “Positioning left-hand movement in violin performance: A system and user study of fingering pattern generation,” in Proc. 26th International Conference on Intelligent User Interfaces (IUI), pp. 208 - 212, Apr. 13 - 17, 2021.

[2] Vincent K. M. Cheung, Hsuan-Kai Kao, and Li Su, “Semi-supervised violin fingering generation using variational autoencoders,” in Proc. International Society for Music Information Retrieval Conference (ISMIR), pp. 113 - 120, Nov. 7 - 12, 2021.

[3] Wei-Yang Lin, Yu-Chiang Frank Wang, and Li Su, “Enhancing violin fingering generation through audio-symbolic fusion,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 811 - 815, Apr. 14 - 19, 2024.

[4] Jie Hu, Li Shen, and Gang Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132 - 7141, Jun. 18 - 22, 2018.

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 2980 - 2988, Oct. 22 - 29, 2017.