한국음향학회지 제43권 제2호 pp. 207~213 (2024)
The Journal of the Acoustical Society of Korea Vol.43, No.2 (2024)
https://doi.org/10.7776/ASK.2024.43.2.207

pISSN : 1225-4428
eISSN : 2287-3775

음고 개수 정보 활용을 통한 기계학습 기반 자동악보전사 모델의 성능 개선 연구

A study on improving the performance of the machine-learning based automatic music transcription model by utilizing pitch number information

이대호,[1] 이석진[1†]

(Daeho Lee[1] and Seokjin Lee[1†])

[1]경북대학교 전자전기공학부

(Received January 23, 2024; accepted February 16, 2024)

초 록: 본 논문은 기계학습 기반 자동악보전사 모델의 입력에 음악적인 정보를 추가하는 방법을 통해 원하는 성능 향상을 얻는 방법을 다루었다. 여기서, 추가한 음악적인 정보는 각 시간 단위마다 발생하는 음고 개수 정보이며, 이는 정답지에서 활성화되는 음고 개수를 세는 방법으로 획득한다. 획득한 음고 개수 정보는 기존 모델의 입력인 로그 멜-스펙트로그램 아래에 연결하여 사용했다. 본 연구에서는 네 가지 음악 정보를 예측하는 네 종류의 블록이 포함된 자동악보전사 모델을 사용하였으며, 각 블록이 예측해야하는 음악 정보에 해당하는 음고 개수 정보를 기존의 입력에 추가해 주는 간단한 방법이 모델의 학습에 도움이 됨을 확인했다. 성능 개선을 검증하기 위하여 MIDI Aligned Piano Sounds (MAPS) 데이터를 활용하여 실험을 진행하였으며, 그 결과 모든 음고 개수 정보를 활용할 경우 프레임 기준F1 점수에서 9.7 %, 끝점을 포함한 노트 기준 F1 점수에서 21.8 %의 성능 향상을 확인하였다.
핵심용어: 자동악보전사, 기계학습, 음고 개수 정보, 다성전사

ABSTRACT: In this paper, we study how to improve the performance of a machine learning-based automatic music transcription model by adding musical information to the input data. Where, the added musical information is information on the number of pitches that occur in each time frame, and which is obtained by counting the number of notes activated in the answer sheet. The obtained information on the number of pitches was used by concatenating it to the log mel-spectrogram, which is the input of the existing model. In this study, we use the automatic music transcription model included the four types of block predicting four types of musical information, we demonstrate that a simple method of adding pitch number information corresponding to the music information to be predicted by each block to the existing input was helpful in training the model. In order to evaluate the performance improvement proceed with an experiment using MIDI Aligned Piano Sounds (MAPS) data, as a result, when using all pitch number information, performance improvement was confirmed by 9.7 % in frame-based F1 score and 21.8 % in note-based F1 score including offset.
Keywords: Automatic music transcription, Machine learning, Pitch number information, Polyphonic traanscription

PACS numbers: 43.10.Vx, 43.50.Ed

†**Corresponding author:** Seokjin Lee (sjlee6@knu.ac.kr)
School of Electronics Engineering, Kyungpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, Republic of Korea
(Tel: 82-53-950-5523, Fax: 82-53-950-5505)

I. 서 론

자동악보전사는 원시 오디오 입력을 자동적으로 악보와 같은 음성 기호로 변환하여 표기해주는 기법이다. 원시 오디오를 음성 기호로 변환하기 위하여 필수적으로 필요한 정보는 음고, 음고의 시작점(onset)과 끝점(offset)이다. 또한, 시작점과 끝점 사이의 음고가 지속되는 부분인 프레임과 건반을 누르는 속도인 벨로시티를 추가해주면 조금 더 자연스러운 소리로 악보를 변환할 수 있다.

Peeling *et al*.[1]은 오디오 신호를 베이지안 확률 모델을 기반으로 해석하여 고전적인 자동악보전사 기법을 제안했다. 오디오 신호는 스펙트로그램으로 변환한 후 행렬 분해(Matrix Factorization)를 이용해 음고와 벨로시티의 곱으로 표현 가능하며, Peeling *et al*.[1]은 기댓값 최대화 알고리즘을 사용해 획득한 확률 변수를 이용하여 분해된 음고와 벨로시티를 예측했다. 전통적인 신호처리 기법을 통한 방법을 기반으로 한 자동악보전사 기법은 Su와 Yang[2]이 소개했다. 한 시점에 동시에 발생하는 음고를 추정하기 위해 스펙트럼과 캡스트럼을 비교하여 기본 주파수와 기본 주기가 공존할 경우 음고의 활성화 여부를 판단했다. Vincent *et al*.[3]은 비음수 행렬 분해(Non-negative Matrix Factorization, NMF) 방법을 이용한 자동악보전사를 연구하였다. 기존의 비음수 행렬 분해를 통한 자동악보전사에서는 추정된 기본 스펙트럼이 불확실하다는 문제가 존재하였으나, 각 기본 스펙트럼을 협대역 스펙트라의 가중 합으로 표현하여 개선했다.

자동악보전사 기술은 전통적인 방식을 통해 해결하기에 어려운 문제였으며, 실제로 숙련된 작곡가나 연주자들에게도 어려운 문제였다. 그렇기에, 최근 기계학습을 이용한 자동악보전사에 대한 연구가 전통적인 방법에 비해 활발히 이루어지고 있다. 기계학습 기반 자동악보전사에서 처음으로 성공적인 결과를 Böck과 Schedl[4]가 보여주었으며, 입력으로 시간과 주파수 각각의 분해능에 중점을 두기 위하여 다른 윈도우 크기로 변환한 두 개의 스펙트로그램을 사용했다. 또한, 양방향 장단기 메모리(Bidirectional Long Short-Term Memory, Bi-LSTM)를 사용하여 시계열 데이터인 오디오 신호의 시계열 정보를 분석할 수 있었다. Sigtia *et al*.[5]는 종단간 구조(end-to-end architecture)를 기반으로 하는 자동악보전사를 위해 음향과 더불어 언어 모델을 함께 학습할 수 있는 방법을 보여줬다. 또한 심층 신경망(Deep Neural Network, DNN), 순환 신경망(Recurrent Neural Network, RNN), 합성곱 신경망(Convolutional Neural Network, CNN)을 이용해 각 신경망에 대한 자동악보전사의 결과를 분석하였다. 합성곱 신경망은 이미지를 처리하는데 두각을 보인 신경망 구조이기 때문에, 단독으로 사용할 경우 자동악보전사에 완벽히 부합하는 네트워크는 아니며, 순환 신경망은 시계열 데이터를 분석하는데 용이한 네트워크로 음향 데이터 분석에는 적합하나 뛰어난 성능을 보여주진 못했다. Hawthorne *et al*.[6]은 이러한 문제점들을 보안하기 위해 합성곱 신경망과 순환 신경망을 연결하여 기계학습 기반 자동악보전사 모델을 구성했다. 해당 네트워크는 다른 구조와 비교했을 때, 월등한 성능을 보여주었으며, 특히 시작점과 프레임을 추정하는 성능이 뛰어났다. 하지만, 끝점 추정에 대한 성능이 여전히 부족하다는 문제점이 존재하며, 이러한 문제점은 끝점이 발생하는 시간대와 비슷한 시간대에 발생하는 프레임의 추정 또한 불안정하게 만들어 프레임 예측 성능을 저하시키는 요소가 된다.

앞선 연구들에서는 기계학습 기반 자동악보전사 모델의 성능을 평가할 때, 주로 프레임 성능에 주목하고 프레임 성능에 비해 다소 낮은 끝점 성능은 간과하였다. 하지만, 프레임 성능을 향상시키기 위해서는 끝점 성능 향상이 필수적이며, 이를 위한 방법에 대한 연구가 필요하다. 기계학습 모델은 인간의 뇌가 학습하는 방식을 본떠 만들어졌기 때문에, 인간이 판단할 때 유용한 정보는 기계학습 모델의 학습에도 도움을 줄 것으로 판단된다. 만일, 사람에게 여러 음이 섞여 있는 화음을 들려준 후 섞여 있는 음의 음고를 분리하라는 문제를 준다면, 아무 정보가 없을 때보다 몇 개의 음이 섞여 있는지 알고 있을 때 더 쉽게 문제를 해결할 수 있다. 이와 관련하여, 비음수 행렬 분해를 기반으로 한 자동악보전사에 대한 Smaragdis와 Brown[7]의 논문에서 관련된 내용에 대한 근거가 존재한다. 비음수 행렬 분해 기법에서 비음수 행렬을 두 개의 비음수 행렬로 분해할 때, 기저

벡터는 분해되는 행렬의 크기를 결정한다. 자동악보전사에 비음수 행렬 분해를 적용할 경우, 스펙트로그램이 비음수 행렬이라고 가정한다면 스펙트로그램을 각각 시간과 주파수에 관한 비음수 행렬로 분해할 수 있다. 이 때, 분해되는 행렬은 기저 벡터의 크기만큼 주파수 조합의 개수를 나타낼 수 있다. 즉, 기저 벡터가 3이라면 분해된 비음수 행렬은 세 가지 조합의 단일 주파수나 주파수 합을 표현할 수 있다. 또한, 기저 벡터는 비음수 행렬 분해 기반 자동악보 전사에서 알고리즘의 성능에 다음과 같은 영향을 준다. 기저 벡터가 표현될 주파수 조합의 개수와 같을 때, 비음수 행렬 분해 기반 자동악보전사 모델은 최고의 성능을 보여준다. 하지만, 기저 벡터와 표현될 주파수 조합의 개수와 다른 경우, 모델이 음고를 추정할 때 오류가 더 많이 발생한다. 이러한 기저 벡터에 대한 연구는 Lee[8]가 진행하였으며, 비음수 행렬 분해 기반 자동악보전사 모델의 기저 벡터를 추정하는 연구를 진행하였다. 이러한 기저 벡터 추정 연구는 음고 개수 정보와 유사한 정보를 포함하고 있다.

본 논문에서는 기계학습 기반 자동악보전사 모델의 끝점 추정 성능 향상을 위해 자동악보전사 모델 입력으로 원시 오디오를 변환한 로그 멜-스펙트로그램에 음고 개수 정보를 추가하여 사용했다. 기존 입력에 음고 개수 정보를 추가하는 단순한 방법만으로도 모델의 성능 향상을 확인할 수 있었다. 논문의 구성은 다음과 같다. II장에는 기준 모델에 대한 설명을 서술하였고, III장에 제안하는 방법인 음고 개수 정보에 대한 설명과 이를 추가하는 방법에 대하여 서술하였으며, IV장에는 자세한 실험 및 평가 방법을 나타내었고, V장에서는 결론 및 향후 연구 방향에 대해 논의하였다.

II. 기계학습 기반 자동악보전사

본 연구에서 기준 모델로 사용한 자동악보전사 모델은 Hawthorne *et al.*[6]의 onsets and frames이며, 이는 기계 학습 기반 자동악보전사 모델 중 성능이 우수한 모델로 알려져 있다. 기준 모델의 구조는 합성곱 신경망(Convolutional Neural Network, CNN), 완전 연결 계층(Fully Connected Layer, FC layer)과 양방향 장

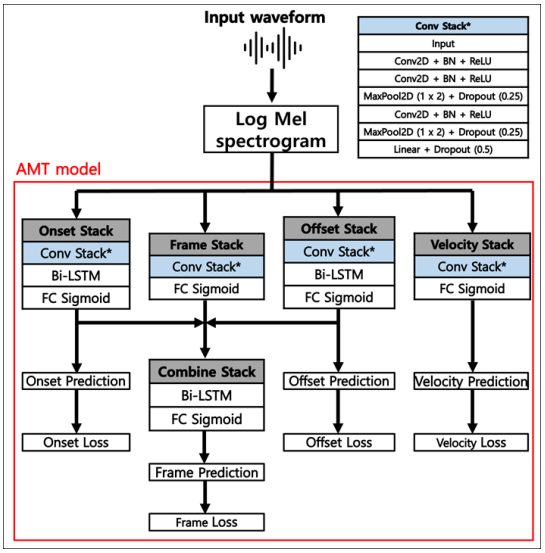

Fig. 1. (Color available online) Diagram of baseline model architecture.[6]

단기 메모리를 연결하여 구성되어 있다. 기준 모델은 총 네 가지 정보를 예측하며, 이는 시작점, 프레임, 끝점, 벨로시티이다. 기준 모델은 각 시간 단위마다 88개의 음고 중에서 활성화되는 음고를 예측하며, $\hat{y}_{onset}, \hat{y}_{frame}, \hat{y}_{offset}, \hat{y}_{velocity} \in \mathbb{R}^{T\times 88}$ 형태로 표현된다. 여기서 T는 멜-스펙트로그램으로 변환된 신호의 시간 단위의 수를 나타내며, 88개의 음고는 피아노 건반의 음고 개수를 기준으로 설정하였다. 기준 모델은 Fig. 1과 같이 네 가지 정보를 예측하기 위해 네 개의 블록으로 구성되어 있다. 프레임 정보는 네 가지 정보 중에서 다른 정보에 비해 많은 비중을 차지하여 더 중요한 정보라고 판단되어, 프레임에 관한 블록은 더 정확한 프레임 예측을 위해 시작점과 끝점의 예측 정보를 사용하도록 구성되어 있다. 기준 모델의 입력은 원시 오디오를 로그 멜-스펙트로그램 형태로 변환하여 사용하며, 이를 각 블록의 입력으로 넣어준다. 여기서 멜-스펙트로그램은 229개의 로그 간격 주파수 빈, 512의 홉 길이, 2048의 윈도우 길이를 사용하여 변환되었으며, 변환된 신호의 크기는 $x_{mel} \in \mathbb{R}^{T\times 229}$ 와 같다.

기준 모델은 프레임의 정확한 예측을 위한 모델 구성을 하였기 때문에, 프레임에 대한 뛰어난 예측 성능을 보여주었으나 끝점 예측은 상반된 결과를 보여주었다. 기준 모델을 통해 예측된 정보들을 살펴

보면 대부분의 오류는 끝점에 해당되는 부분에서 발생하며, 이를 개선한다면 프레임 성능 개선까지 이루어질 수 있을 것이라 기대된다.

III. 음고 개수 정보

이때까지 다양한 기계학습 기반 자동악보전사 모델이 만들어져 왔으나, 대부분의 모델의 입력은 원시 오디오나 원시 오디오를 변환한 형태를 사용했다. 원시 오디오에는 음의 높이, 음의 길이, 박자, 리듬, 코드, 악기 등 다양한 음악적인 정보들이 담겨져 있으나, 기계학습 기반 자동악보전사 모델이 입력 신호의 음악적 정보를 효과적으로 활용하지 못한다고 판단된다. 기계학습 기반 자동악보전사 모델이 입력 오디오 신호에서 해석하지 못한 내용을 추가적으로 입력에 더해준다면, 모델의 학습에 도움을 줄 것으로 가정하고 실험을 진행하였다. 본 연구에서는 다양한 음악적인 정보 가운데 앞서 서론에서 언급한 근거를 기반으로 각 시간 단위마다 발생하는 음고 개수 정보를 모델의 입력에 추가해 주었다.

II장에 서술된 것과 같이, 본 연구에서 사용된 기준 모델은 서로 다른 네 가지 정보를 예측해야하는 네 가지 블록으로 구성되어 있으나, 모든 블록의 입력은 공통적으로 로그 멜-스펙트로그램을 사용하고 있다. 여기에 각 시간 단위마다 발생하는 음고 개수 정보를 추가해주면, 각 블록은 예측해야하는 정보에 더욱 부합하는 입력을 사용할 수 있게 된다. 여기서 음고 개수 정보는 모델 정답지에서 각 시간 단위마다 활성화하는 각 정보의 음고 개수의 합으로 $n_{onset}, n_{frame}, n_{offset} \in \mathbb{Z}^{T\times 1}$ 형태로 표현된다. Fig. 2를 통해 예시를 표현했으며, 위쪽은 정답지 정보로

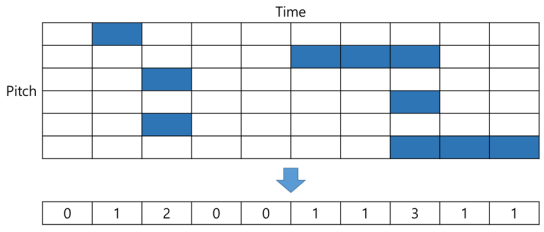

Fig. 2. (Color available online) Example of pitch number information counting.

음고가 활성화된 구간을 파란색으로 표시했고, 아래는 획득한 음고 개수 정보로 각 시간 단위마다 활성화되는 음고의 개수를 세어 정답지의 시간 단위의 개수만큼 표현된다.

본 연구에서 각 시간 단위마다 활성화되는 음고의 최댓값은 20으로 설정하였다. 이는, 실험에 사용된 데이터는 피아노 연주가 녹음된 오디오 신호이기 때문에, 두 사람이 피아노를 연주했다고 가정했을 때 최대로 사용될 수 있는 손가락 개수인 20개로 설정했다. 만약 데이터를 통해 얻은 음고 개수 정보가 20을 초과하는 경우, 20으로 대체하여 사용하였다. 실험에 사용한 MAPS 데이터에서 음고 개수 정보가 20을 초과하는 경우는 0.001 %에 해당되며, 일반적인 음악에서도 동시에 20개 이상의 음을 연주하는 것이 일반적이지 않기 때문에, 위와 같은 방안이 학습에 큰 영향을 끼치지 않을 것으로 판단했다. 집계된 음고 개수 정보는 기존의 모델 입력인 멜-스펙트로그램 아래에 연결하여, $x_{onset}, x_{frame}, x_{offset} \in \mathbb{R}^{T\times 230}$ 형태로 표현된다.

IV. 실 험

본 논문에서는 자동악보전사의 성능 향상을 위해 기준 모델을 수정하여 사용했다. 합성곱 신경망 채널은 48/48/96으로, 장단기 메모리 유닛은 256으로, 완전 연결 유닛은 768로 키워서 사용했다. 학습을 위해 파이토치(pytorch)[9]를 사용했고, 4의 배치 크기, 0.005의 학습률, L2 노름의 임계값이 3인 그래디언트 클리핑을 사용했다. GPU 용량의 한계로 기준 모델을 다룬 Reference [6]에서 언급된 것에 비해 작은 배치 크기를 사용하였으며, 모델의 최적화를 위해 아담(Adam)[10]과 함께 50,000번 반복 학습을 진행하였고, 0.98의 학습률 감소를 10,000번 반복 학습마다 적용하였다.

실험을 위해 사용된 데이터셋은 270곡의 피아노 연주가 녹음된 MAPS 데이터셋이다.[11] 학습 데이터는 150곡의 연주를 사용하였고, 평가와 검증을 위해 각 60곡의 연주를 사용했다. 학습에 사용된 모든 오디오 신호는 16 kHz로 다운샘플링하였으며, 무작위로 20 s를 잘라 사용했다.

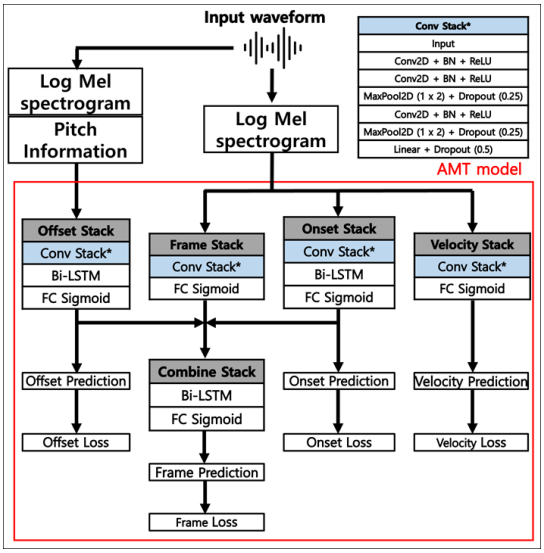

Fig. 3. (Color available online) Diagram of proposed network architecture with offset numbers of pitch information.

총 5종류의 실험을 진행하였으며, 시작점, 프레임, 끝점의 음고 개수 정보를 하나씩 사용한 실험, 끝점 음고 개수 정보를 로그 멜-스펙트로그램 위에 합쳐 입력으로 사용한 실험, 그리고 마지막으로 세 종류의 음고 개수 정보를 모두 사용한 실험을 진행하였다. 끝점 음고 개수 정보를 로그 멜-스펙트로그램 위에 합쳐 입력으로 사용한 실험을 제외한 실험들에서는 음고 개수 정보를 로그 멜-스펙트로그램 아래에 붙여 입력으로 사용했다. 또한, 음고 개수 정보를 하나만 사용한 실험은 음고 개수 정보와 동일한 블록에만 음고 개수 정보를 포함한 입력이 사용되고 다른 블록에는 기존 입력인 로그-멜스펙트로그램을 사용하였다. 예를 들어, 끝점 음고 개수 정보를 추가한 실험에서 끝점 블록의 입력으로 로그 멜-스펙트로그램 아래에 끝점 음고 개수 정보를 추가한 데이터를 사용하였고, 나머지 블록은 로그 멜-스펙트로그램이 입력으로 사용되었다. Fig. 3에 끝점 음고 개수 정보를 추가한 학습 방법을 표현했다.

각 실험을 통해 학습된 자동악보전사 모델을 이용해 예측한 피아노 연주는 Fig. 4에 표현했다. Fig. 4의 ground truth 는 정답을, baseline 은 기준 모델의 예측 결과를, 그리고 나머지는 제안하는 바와 같이 음고 개수 정보를 추가하여 학습한 모델의 예측 결과들을 보여준다. 기준 모델은 없는 음고를 있다고 판단한

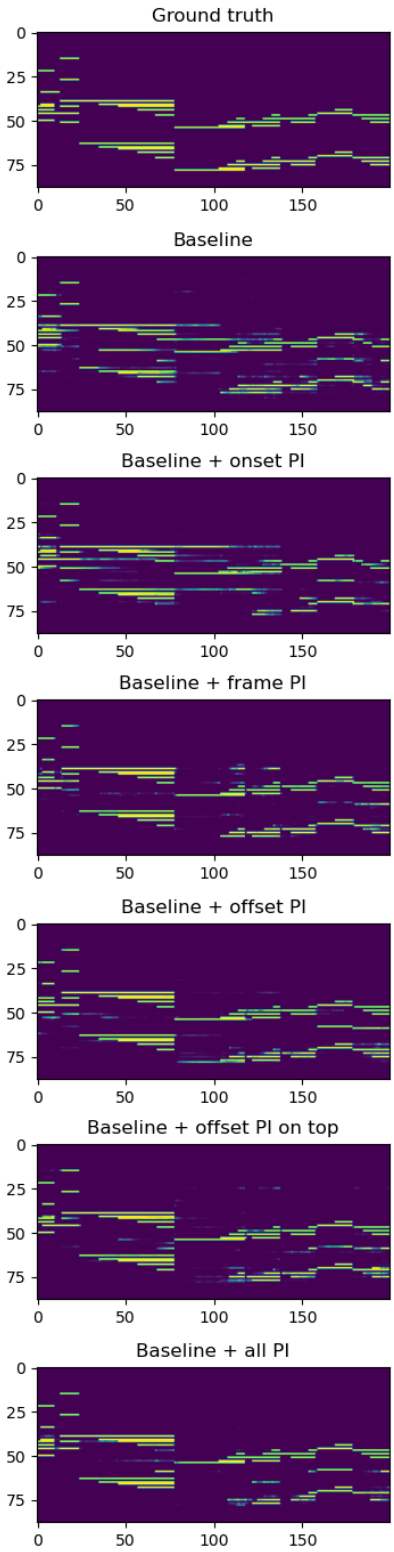

Fig. 4. (Color available online) Transcription results of baseline and propsed models. X-and y-axes denote the frame number and MIDI note number, respectively.

Table 1. F1-scores of each model. P, R, F1 denotes the precision, recall and F1 score, respectively.

	Frame			Note			Note with offset			Note with offset & velocity		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	82.1	78.5	79.8	93.7	88.0	90.6	59.6	56.2	57.8	57.0	53.7	55.2
Baseline + onset pitch information	82.3	78.9	80.3	93.5	**89.6**	**91.5**	57.4	55.2	56.3	54.9	52.8	53.8
Baseline + frame pitch information	**93.2**	77.4	84.3	**94.9**	86.2	90.2	77.8	70.8	74.0	74.4	67.8	70.8
Baseline + offset pitch information	90.6	**82.2**	**86.1**	92.9	88.0	90.2	78.1	**74.1**	**75.9**	74.6	**70.8**	72.6
Baseline + offset pitch information on top	92.1	80.7	85.9	93.6	87.4	90.2	**78.3**	73.3	75.6	**75.3**	70.5	**72.7**
Baseline + all pitches information	**95.7**	**84.3**	**89.5**	**94.0**	**90.9**	**92.4**	**80.9**	**78.4**	**79.6**	**78.3**	**75.9**	**77.0**

경우도 많으며, 끝점을 정확하게 판단하지 못하는 결과를 보여준다. 시작점 음고 개수 정보를 추가한 경우도 마찬가지로 끝점을 정확하게 판단하지 못하는 결과를 보여준다. 프레임, 끝점 음고 개수 정보를 추가하여 학습한 각 모델은 모두 기준 모델에 비해 끝점을 정확하게 예측하는 결과를 얻었다.

실험의 평가는 F1 점수를 통해 진행하였으며, 각기 다른 총 네 가지 기준으로 평가하여 Table 1에 나타냈다. 여기서 note 기준은 시작점에 대한 평가를, frame 기준은 프레임에 대한 평가를, note with offset은 시작점과 끝점에 대한 평가를, note with offset & velocity는 시작점, 끝점, 벨로시티에 대한 평가를 진행한 결과이며, 정답위치에서 ± 50 ms 범위까지 오차범위를 허용하였다. Baseline은 기준 모델을 통해 얻은 결과이며, 나머지 항목들은 각각의 음고 개수 정보를 추가하여 실험한 모델의 결과이다. 결과 중에서 가장 높은 점수는 진한 글씨와 함께 밑줄을 그어 표현하였고, 두 번째로 높은 점수를 획득한 값은 진한 글씨로 표현하였다. 기준 모델의 결과를 보면, 노트 기준 점수는 프레임 기준 점수에 비해 높으며 끝점과 연계된 노트 기준 점수는 낮은 성능을 기록한 것을 확인할 수 있다. 시작점 음고 개수 정보를 추가한 경우 노트와 프레임 기준 F1 점수가 모두 향상된 것을 확인할 수 있으나, 끝점 관련 노트 기준 성능은 오히려 하락하였다. 시작점 음고 개수 정보는 끝점과 연관되지 않은 정보이기 때문에, 시작점과 연관된 노트와 프레임 기준 점수는 향상시키고 끝점 관련 점수는 큰 영향을 주지 않음을 확인할 수 있는 결과이다. 프레임 음고 정보 개수를 추가한 경우, 노트 기준 F1 점수는 기준 모델의 성능에 비해 0.4 % 하락하였으나, 프레임 기준 F1 점수는 6.3 % 향상하였으며, 끝점과 연관된 노트 기준 점수도 전반적으로 상승하였다. 끝점 음고 개수 정보를 추가한 경우도 마찬가지로 프레임 음고 개수 정보를 추가한 경우과 유사한 결과를 획득하였으며, 끝점 연관 노트 점수는 기준 모델의 결과에 비해 18.2 % 향상된 점수를 프레임 기준 F1 점수는 6.3 % 향상된 점수를 획득하였다. 끝점 음고 개수 정보를 로그 멜-스펙트로그램의 위에 붙여 입력으로 사용한 경우는, 끝점 음고 개수 정보를 로그 멜-스펙트로그램 아래에 붙여 입력으로 사용한 경우와 유사한 성능을 획득하였으며, 두 방법 사이에 큰 결과 차이가 없음을 확인하였다. 마지막으로 모든 음고 개수 정보를 추가한 경우, 가장 높은 점수를 획득하였으며, 프레임 기준 F1 점수가 노트 기준 F1 점수에 비해 차이가 컸던 다른 실험과 달리 두 기준의 성능 차를 2.9 %까지 줄인 결과를 얻었다. 실험을 통해 음고 개수 정보를 로그 멜-스펙트로그램에 붙여 입력으로 사용하는 변화만으로도 기존 결과에 비해 전반적으로 향상된 성능을 획득할 수 있음을 F1 점수를 통해 확인 할 수 있었다. 특히 본 연구의 목표인, 끝점과 관련된 음악적인 정보를 추가하여 향상된 끝점 성능이 프레임 예측에도 도움이 될 수 있음을 확인하였다.

V. 결 론

기계학습 기반 자동악보전사 모델은 시작점과 프레임에 대한 예측은 높은 성능을 보여주나, 끝점에 대한 예측은 낮은 성능을 보여주는 문제가 있었다. 본 논문에서는 이러한 문제를 해결하기 위해 각 시

간 단위마다 음고 개수를 세어 획득한 음고 개수 정보를 기존의 모델 입력에 연결하여 사용하는 방법을 제안했다.

같은 모델 구조를 사용하지만 간단히 기존 입력에 음악적인 정보를 추가하는 방법만으로도 기존 모델에 비해 향상된 성능을 획득할 수 있었다. 끝점 음고 개수 정보를 추가한 경우, 끝점을 포함한 노트 기준 F1 점수는 18.1 % 향상되었으며, 이로 인해 프레임 기준 F1 점수 또한 6.3 % 향상되었다. 모든 음고 개수 정보를 추가한 경우 모든 실험 중 가장 높은 점수를 획득하였으며, 끝점을 포함한 노트 기준 F1 점수에서 21.8 %, 프레임 기준 F1 점수에서 9.7 %의 성능 향상을 보여주었다.

음고 개수 정보를 이용하기 위해서는 모델의 입력으로 사용되는 오디오 신호의 정답지가 필요하다는 한계점이 있으나, 음고 개수 정보를 이용할 경우 자동악보전사 모델의 최대 성능 향상을 확인할 수 있었다. 이후 추가적인 연구 방향으로써, 음고 개수 정보를 추정하는 모델의 개발을 통해 위의 문제를 해결하고자 한다. 또한, 음고 개수 정보가 아닌 다른 음악적인 정보의 추가를 통해 원하는 성능 향상을 할 수 있을 것이라 기대된다.

References

1. P. H. Peeling, A. T. Cemgil, and S. J. Godsill, “Generative spectrogram factorization models for polyphonic piano transcription,” IEEE Trans. on Audio, Speech, and Lang. Process. **18**, 519-527 (2009).
2. L. Su and Y.-H. Yang, “Combining spectral and temporal representations for multipitch estimation of polyphonic music,” IEEE/ACM Trans. on Audio, Speech, and Lang. Process. **23**, 1600-1612 (2015).
3. E. Vincent, N. Bertin, and R. Badeau, “Adaptive harmonic spectral decomposition for multiple pitch estimation,” IEEE Trans. on Audio, Speech, and Lang. Process. **18**, 528-537 (2009).
4. S. Böck and M. Schedl, “Polyphonic piano note transcription with recurrent neural networks,” Proc. IEEE ICASSP, 121-124 (2012).
5. S. Sigtia, E. Benetos, and S. Dixon, “An end-to-end neural network for polyphonic piano music transcription,” IEEE/ACM Trans. on Audio, Speech, and Lang. Process. **24**, 927-939 (2016).
6. C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, “Onsets and frames: Dual-objective piano transcription,” arXiv preprint arXiv:1710.11153 (2017).
7. P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” Proc. IEEE WASPAA, No. 03TH8684 (2003).
8. S. Lee, “Estimating the rank of a nonnegative matrix factorization model for automatic music transcription based on stein’s unbiased risk estimator,” Appl. Sci. **10**, 2911 (2020).
9. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” Proc. NeurIPS, 1-12 (2019).
10. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980 (2014).
11. V. Emiya, N. Bertin, B. David, and R. Badeau, “MAPS-A piano database for multipitch estimation and automatic transcription of music,” INRIA, Research Rep., 2010.

저자 약력

▸ 이 대 호 (Daeho Lee)

2020년 2월 : 동아대학교 전자공학부 학사
2022년 8월 : 경북대학교 전자전기공학부 석사
2022년 3월 ~ 현재 : 경북대학교 전자전기공학부 박사과정

▸ 이 석 진 (Seokjin Lee)

2006년 8월 : 서울대학교 전기컴퓨터공학부 학사
2008년 8월 : 서울대학교 전기컴퓨터공학부 석사
2012년 2월 : 서울대학교 전기컴퓨터공학부 박사
2012년 3월 : ㈜LG전자 CTO연구소 선임연구원
2014년 3월 : 경기대학교 전자공학과 조교수
2018년 3월 : 경북대학교 전자공학부 조교수
2020년 10월 ~ 현재 : 경북대학교 전자공학부 부교수