End-to-end Automatic Music Transcription of Polyphonic Music Using Convolutional Neural Networks

2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS) | 979-8-3315-4010-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISAS64331.2024.10845524

Emin Germen
Eskisehir Technical University
Electrical & Electronics Eng. Dept
Eskisehir, Turkey
0000-0003-1301-3786

Can Karadoğan
Istanbul Technical University
State Conservatory MIAM
Istanbul, Turkey
0000-0003-3611-6980

***Abstract*—This paper presents an automatic music transcription model based on Convolutional Neural Networks (CNNs) that mimics the "trained ear" in music recognition. The approach pushes forward the fields of signal processing and music technology, with a focus on multi-instrument transcription featuring traditional Turkish instruments like the Qanun and Oud, known for their distinct timbral qualities and early decay characteristics. The study involves creating multi-pitch datasets from very basic combinations, training the CNN on this data, and achieving high transcription accuracy considering the F1 scores for two-part compositions. The training process equips the model to understand the fundamental traits of individual instruments, enabling it to identify and separate complex patterns in mixed audio. The aim is to enhance the model's ability to distinguish and analyze specific musical elements, supporting applications in music production, audio engineering, and music education.**

Keywords—Music transcription, constant Q transform, convolutional neural network, signal processing

I. Introduction

The Automated Music Transcription (AMT) deals with recognizing musical signals and converting them into musical notation or MIDI data.[1] When music is played with a monophonic instrument, automatically transcribing it into sheet music is not a particularly challenging problem. However, transcribing music that involves multiple instruments or polyphonic instruments like the piano, organ, or guitar becomes a much more complex task due to the need to accurately distinguish overlapping pitches from different instruments, which often interact in complex ways. Various disciplines have contributed solutions to this problem, each offering approaches from their unique perspectives. At the foundation of solving this problem, Blind Source Separation (BSS) techniques have played a crucial role.[2] Early approaches relied on linear system models to separate individual sound sources. Among these methods, Nonnegative Matrix Factorization (NMF) [3] has been widely explored as a solution for decomposing complex audio signals into their constituent components. NMF works by factorizing a matrix representation of the audio signal into nonnegative factors, corresponding to the underlying sources and their contributions over time, making it particularly useful for distinguishing different instruments or sound elements in a mixed audio environment.[4]

Automated music transcription is evolving towards end-to-end models that aim to streamline the transcription process from audio input to musical notation output. Advancements in machine learning and deep neural networks, along with the development of sophisticated models, are leading to increasingly competent results in both recognizing and separating instruments in music featuring multiple instruments, as well as transcribing polyphonic music into notation.[5], [6], [7] These innovations are pushing the boundaries of what is achievable in automated music analysis, however, we are quite far from achieving 100% solutions.[8]

The literature on transcribing polyphonic music with multiple instruments predominantly relies on corpora developed for Western classical music and its associated instruments. As a result, researches focused on analyzing traditional instruments remains limited.[9] In this study, we address this gap by creating a simple corpus for polyphonic music featuring traditional Turkish musical instruments, namely the stringed instruments *ud* and *qanun*. Despite the simplicity of the corpus, we were able to design an end-to-end automated music transcription model that achieved promising results. This model demonstrates that even with a basic dataset, it is possible to obtain a reasonable level of accuracy in transcribing music involving traditional instruments, highlighting the potential for further development in this area. A deep Convolutional Neural Network (CNN) [10], [11] model has been developed and trained using CNN corpus obtained by introducing the basic combinations of the two instruments which describe the fundamental characteristics of the amalgamated sounds from these instruments. The trained model is then employed to distinguish polyphonic music compositions involving these instruments. The outcomes of this approach have been notably satisfactory.

The paper's organization is structured as follows: Chapter Two presents a comprehensive description of the model utilized for transcribing two-part music. This chapter will detail the preparation of the dataset used for training, as well as the features involved. Chapter Three is devoted to an in-depth discussion of the results derived from this study. Conclusively, Chapter Four offers a summation and final thoughts of the paper.

II. Automatic Music Transcription Model

In this section, the designed end-to-end model will be described in detail. The model, which is fundamentally based on a CNN architecture, will be explained, including how the dataset was prepared and how the model was trained using

979-8-3315-4010-4/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on July 03,2025 at 15:25:45 UTC from IEEE Xplore. Restrictions apply.

this dataset. In addition to these aspects, the discussion will cover how the model's outputs align with the results of onset music analysis, which is used to characterize the musical structure. The architecture of the designed model is presented in **Error! Reference source not found.** 1, providing a visual reference for the discussed components.

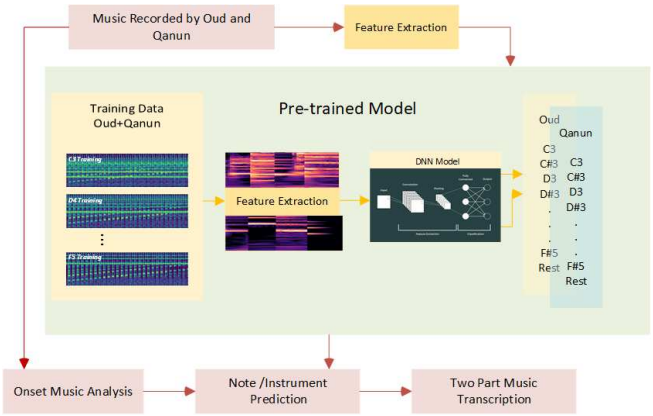

Fig. 1. The End-to-end Music Transcription system

A. Data Corpus to train the CNN Model

The main aim of this model is to better understand ear training by examining fundamental instrumental combinations in music. The research focuses on creating a multi-instrument, multi-pitch dataset using music that blends Turkish Qanun and Oud. A key part of the process involves separating the musical pieces to capture the unique characteristics of the two instruments when played together.

Here, a "well-trained ear" means the ability to accurately recognize and interpret sounds, especially in music. This skill is highly valued among musicians, audio engineers, and music enthusiasts, with pitch recognition in polyphonic music being a crucial element. The study aims to see if this kind of training can be replicated by using note combinations from two instruments played simultaneously.

The approach is straightforward: the Oud plays a single note (e.g., C3), while the Qanun plays various notes across three octaves, recording all possible combinations. Fig 2 shows an example of the training data for the A4 note on the Qanun, where the Qanun plays A4 while the Oud chromatically moves from C3 to A5. While the method can handle microtonal intervals, they are not included in this study. Each note's training sample lasts 56 seconds, with A4 as the predicted output.

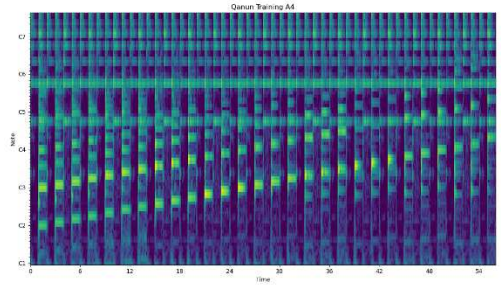

Fig. 2. Qanun data samples for Note A4

Altogether, 37 distinct datasets were created for the Qanun, covering chromatic notes from C3 to A5, resulting in a total of 2,072 seconds of data. An additional dataset was added to represent pauses, where only the Oud plays or there's silence, making 37 datasets in total for recognizing Qanun phrases, including C3, C#3, D3, D#3, … A5, and Rest.

The same process was used for the Oud. This systematic method ensures a broad coverage of the musical range, capturing pitch variations and the harmonic relationship between the instruments. By focusing on note combinations, the study aims to better understand and model the complexities of musical interactions.

The dataset for both the Qanun and Oud was prepared for CNN training by generating spectrograms using the Constant Q Transform (CQT). [12] CQT bins were configured to cover a range of 62 notes, ensuring comprehensive pitch representation. The spectrograms were constructed with a sliding window of 512 points, offering a time resolution of 23 milliseconds, which allows for precise detection of rapid changes in the musical lements, which involves capturing the distinct multi-timbral characteristics of these two instruments when played together. The CQT intervals were defined according to the 12 fundamental semitones that form the basis of modern Western harmonic music as shown in Equation 1 and 2. However, to better accommodate maqam music, the full scale could be subdivided into finer intervals to create a specialized corpus.

$$f_m = f_{min} 2^{\frac{m}{12*b}} \quad (1)$$

Here f_{min} is the minimal frequency in analysis, and b is the number of bins per semi-tone. The CQT can be derived with a quality factor $Q = \frac{f_m}{f_{m+1} - f_m} = 1/(2^{\frac{1}{12b}} - 1)$

$$F_{CQT}[m, \Lambda] = \frac{1}{N_m} \sum_{k=0}^{N_m - 1} x[k + \Lambda R] w_m[k] e^{\left(-i\frac{2\pi Q k}{N_m}\right)} \quad (2)$$

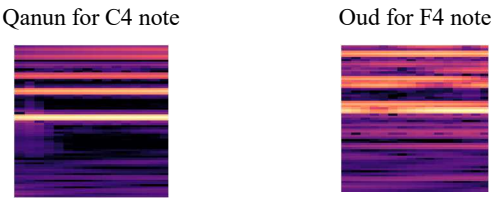

Fig. 3. Examples of corpus for Qanun and Oud

Each resulting 52-second spectrogram was divided into 301 equal segments, from which training figures were generated, as shown in Fig. 3. In addition to these training data, the transient moments of the instruments were captured using onset analysis, and CQT spectrograms corresponding to these exact onset points were also included in the corpus. At the end, 12350 images for Qanun and 12200 images for Oud have been obtained.

B. CNN Model

Several models were developed and rigorously tested, with extensive experimentation and hyperparameter tuning conducted to optimize performance. The final architecture, shown in Fig. 4, was selected based on its superior results and was subsequently trained using the instrument corpus. The model generates outputs by predicting both the instrument and the corresponding musical note, allowing it to classify 32 distinct notes in addition to handling the rest which means the instrument is not played.

Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on July 03,2025 at 15:25:45 UTC from IEEE Xplore. Restrictions apply.

To evaluate the model's performance, the dataset was divided into a 75% training set and a 25% validation set, ensuring a balanced approach to training and validation. The training process involved 200 epochs, utilizing an i9 processor and an RTX 4070 GPU for computational efficiency. This configuration enabled the model to complete the training in 11 hours and 34 minutes, demonstrating the practicality of the setup for large-scale corpus training. In the end, the obtained training accuracy was %93.23 and the validation was %91.04.

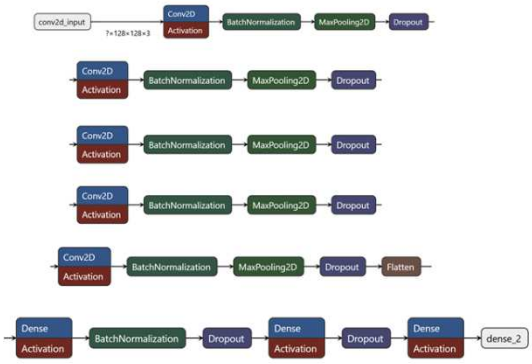

Fig. 4 CNN Model

C. *Transient Detection and Onset Analysis*

Music transcription entails detecting temporal changes in a piece, such as shifts in the notes played or moments of silence. These fluctuations, known as transients, are most prominent during the initial attack phase of a note. In this study, which centers on plucked instruments like the Oud and Qanun, such transients are distinctly observable. To accurately detect these transients and identify structural changes in the musical time series, both the energy envelope and zero crossings are employed. Additionally, frequency-domain features such as spectral centroid, spectral skewness, and spectral kurtosis are used to pinpoint potential onset locations more precisely.

Fig. 5 and Fig. 6 illustrate this process, displaying examples of onset detection for the Oud and Qanun in both time and frequency domains. This multifaceted approach captures the intricate nuances of the instruments' interplay, enabling precise transcription of the music by accounting for the dynamic variations inherent to plucked string instruments.

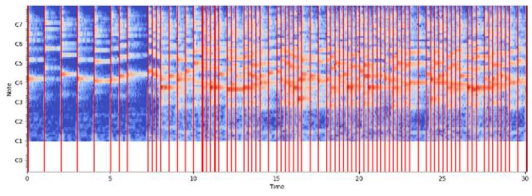

Fig. 5. Onset times of Qanun Oud Music Example on CQT spectrogram

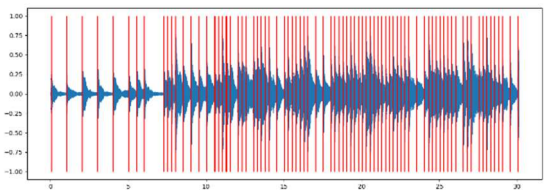

Fig. 6. Onset times of Qanun Oud Music in the time domain

D. *The Transcription of Qanun Oud Music*

For Automatic Music Transcription, a CNN trained over 200 epochs was used. The analysis focused on two-part music featuring the Oud and Qanun. Around the detected onset times, 11 CQT images were generated to identify the potential notes played on either the Oud or Qanun during those transient moments. Specifically, five CQT transform windows, each lasting 23 milliseconds, were shifted at intervals of 16.5 milliseconds both before and after the onset, creating a total of 11 images. In other words, these images cover a time of ±82.5 milliseconds around each onset point.

For each transient marked by the onset time, the classification results from the CNN for both the instrument and the notes played by that instrument were evaluated across the 11 different images. To identify the transient note, the most frequent classification result across the analyzed images is chosen. The results were quite successful when either the Oud or the Qanun was played individually, but ambiguity arose when both instruments were played simultaneously. When the transcription was focused solely on either the Oud or the Qanun, the instrument identified in the classification was considered, and the corresponding musical notes were determined accordingly.

In performing the music transcription, automatic timing adjustments based on onset analysis were not made, and it was assumed that the tempo of the music was known in advance. In the experiments, the music was played at a tempo of 60 beats per minute. Additionally, it was assumed that the smallest note value was a sixteenth note, with the timing structured as combinations of 250-millisecond notes. The primary aim here was to determine whether a CNN model, trained on a simple corpus, could accurately classify notes during transient moments.

III. The Results

The Bach composition "Art of Fugue: Kontrapunctus Nr. 1" was recorded using sampled instruments and analyzed using the proposed framework. The music was sampled at a rate of 44100 Hz, and 62 bins CQT transforms were obtained and converted to image files. The time resolution is 23 milliseconds per window. The test data consisted of 30 seconds of music. For this data set the onset points were tracked, and the duration of single instrument data was extracted by analyzing consecutive points. By knowing the shift in the time axis and sample rate, the note values and durations were determined and converted into MIDI data. The MIDI data was then imported into MuseScore and converted into sheet music notation.

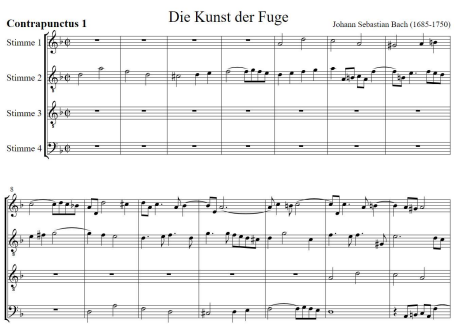

Fig. 7. Bach Contrapuntus to evaluate the framework.

Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on July 03,2025 at 15:25:45 UTC from IEEE Xplore. Restrictions apply.

As shown in Fig. 7, the music has been created by Stimme 1 and Stimme 2. Stimme 1 has been played with Oud, and Stimme 2 has been played with the Qanun. After End-to-end music transcription obtained by the proposed system, the extracted Qanun and the Oud parts have been shown in Fig. 8 and Fig. 9.

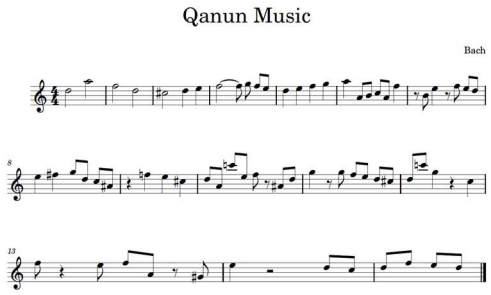

Fig. 8. Transcribed Stimme 2 played by Qanun.

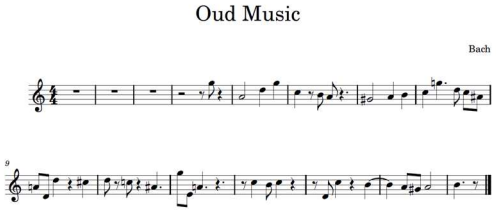

Fig. 9. Transcribed Stimme 1 played by Oud.

To evaluate the results, a method was used in which a prediction is considered a true positive only if all relevant attributes are correctly identified. For frame-level predictions, this means accurately predicting both the pitch and instrument class. For note-level predictions, the pitch, onset, and instrument class must all be correct.

Precision, Recall, and F1 score are commonly used metrics for evaluating the performance of a classification model. These metrics are computed using the counts of true positive (TP), false positive (FP), and false negative (FN) predictions made by the model on a test dataset.

Precision is a measure of the accuracy of the model's positive predictions. It is calculated as the ratio of true positive predictions to the total number of positive predictions made by the model:

$$P = TP / (TP + FP)$$

The recall is a measure of the model's ability to identify all positive instances in the test dataset correctly. It is calculated as the ratio of true positive predictions to the total number of actual positive instances in the test dataset:

$$R = TP / (TP + FN)$$

The F1 score is a metric that combines precision and recall into a single score. It is calculated as the harmonic mean of precision and recall:

$$F1 = 2 * (P * R) / (P + R)$$

The F1 score is a valuable metric when it is needed to balance precision and recall, as it gives equal weight to both. However, it is essential to note that optimizing for one of these metrics may come at the expense of the other, so it is often necessary to consider both precision and recall when evaluating the performance of a model.

To calculate the performance metrics, the smallest note value in the music was identified, and all other notes were represented as integer multiples of that value. For example, in the music shown in Fig. 7, the quaver (eight notes) was identified as the smallest note value, and all other notes were described as multiples of the eight notes. The transcribed music and the corresponding MIDI data were compared, and a confusion matrix was calculated based on the individual note values. The results are shown in Table 1 and Table 2.

TABLE I. QANUN TRANSCRIPTION EVALUATION METRICS FOR EACH PREDICTED NOTE.

	Precision	Recall	F1-score
A#3	0.00	0.00	0.00
A3	0.50	1.00	0.67
A4	1.00	1.00	1.00
B3	1.00	1.00	1.00
C#4	0.71	1.00	0.83
C4	0.80	1.00	0.89
C5	0.00	0.00	0.00
D4	0.95	0.87	0.91
E4	1.00	0.72	0.84
F#4	1.00	1.00	1.00
F4	1.00	0.76	0.86
G#3	1.00	1.00	1.00
G4	1.00	0.53	0.69
Rest	0.00	0.00	0.00
Accuracy			**0.78**
Macro avg	**0.71**	**0.71**	**0.69**
Weighted avg	**0.96**	**0.78**	**0.85**

TABLE II. OUD TRANSCRIPTION EVALUATION METRICS FOR EACH PREDICTED NOTE.

	precision	recall	f1-score
A4	1.00	0.50	0.67
B-4	0.00	0.00	0.00
B4	0.89	0.73	0.80
C#5	1.00	1.00	1.00
C5	1.00	0.48	0.65
D4	1.00	1.00	1.00
D5	1.00	0.60	0.75
E4	1.00	1.00	1.00
G#4	1.00	0.60	0.75
G5	0.00	0.00	0.00
Accuracy			**0.66**
Macro avg	**0.70**	**0.57**	**0.61**
Weighted avg	**0.82**	**0.66**	**0.68**

IV. CONCLUSION

This study reveals that even with a limited training dataset, a CNN model trained on CQT spectrogram data from two-part compositions featuring traditional Turkish instruments can achieve impressive results. While the performance is not flawless, the approach serves as an important preliminary step, akin to how learning intervals and tonal nuances form the basis of ear training in humans, helping to distinguish musical patterns.

The current method already works well with complex Western music, but there's room to grow, especially when it comes to the unique structures found in classical Turkish music. This isn't just about pushing tech boundaries; it's about cultural preservation. By improving how we transcribe Maqam and other traditional Turkish forms, we're helping to keep a rich musical history alive. Being able to accurately capture and study these complex sounds not only advances research but

Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on July 03,2025 at 15:25:45 UTC from IEEE Xplore. Restrictions apply.

also brings traditional music closer to the tools and understanding of today's world.

REFERENCES

[1] E. Benetos, S. Cherla, and T. Weyde, "An Efficient Shift-Invariant Model for Polyphonic Music Transcription." [Online]. Available: https://code.soundsoftware.ac.uk/projects/amt_mssiplca_fast

[2] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler, "Automatic Music Transcription and Audio Source Separation," *Cybernetics &Systems*, vol. 33, no. 6, pp. 603–627, Sep. 2002, doi: 10.1080/01969720290040777.

[3] N. Bertin, R. Badeau, and G. Richard, "Blind Signal Decompositions for Automatic Transcription of Polyphonic Music: NMF and K-SVD on the Benchmark," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, IEEE, 2007, pp. I-65-I–68. doi: 10.1109/ICASSP.2007.366617.

[4] S. Ansari *et al.*, "A survey of artificial intelligence approaches in blind source separation," *Neurocomputing*, vol. 561, p. 126895, 2023, doi: https://doi.org/10.1016/j.neucom.2023.126895.

[5] S. Sigtia, E. Benetos, N. Boulanger-Lewandowski, T. Weyde, A. S. d'Avila Garcez, and S. Dixon, "A hybrid recurrent neural network for music transcription," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2015, pp. 2061–2065. doi: 10.1109/ICASSP.2015.7178333.

[6] S. Sigtia, E. Benetos, and S. Dixon, "An End-to-End Neural Network for Polyphonic Piano Music Transcription," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 24, no. 5, pp. 927–939, May 2016, doi: 10.1109/TASLP.2016.2533858.

[7] R. Nishikimi, E. Nakamura, M. Goto, and K. Yoshii, "Audio-to-score singing transcription based on a CRNN-HSMM hybrid model," *APSIPA Trans Signal Inf Process*, vol. 10, no. 1, p. e7, 2021, doi: 10.1017/ATSIP.2021.4.

[8] K. Tanaka, T. Nakatsuka, R. Nishikimi, K. Yoshii, and S. Morishima, "Multi-instrument Music Transcription Based on Deep Spherical Clustering of Spectrograms and Pitchgrams," *ISMIR, International Society for Music Information Retrieval Conference*, pp. 327–334, 2020.

[9] E. Germen and C. Karadoğan, "END-TO-END AUTOMATIC MUSIC TRANSCRIPTION OF POLYPHONIC QANUN AND OUD MUSIC USING DEEP NEURAL NETWORK," *Eskişehir Technical University Journal of Science and Technology A - Applied Sciences and Engineering*, vol. 25, no. 3, pp. 442–455, Sep. 2024, doi: 10.18038/ESTUBTDA.1467350.

[10] R. Venkatesan and B. Li, "Convolutional Neural Networks in Visual Computing: A Concise Guide," *Convolutional Neural Networks in Visual Computing: A Concise Guide*, pp. 1–168, Oct. 2017

[11] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature 2015 521:7553*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[12] J. Brown, "Calculation of a Constant Q Spectral Transform," *Journal of the Acoustical Society of America*, vol. 89, p. 425, Dec. 1991, doi: 10.1121/1.400476.

Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on July 03,2025 at 15:25:45 UTC from IEEE Xplore. Restrictions apply.