# Research for Quantitative Character Analysis in Opera through Markov Chains

Jiaying Gao

The Experimental High School attached to Beijing Normal University, Beijing, 100032, China

gaojiaying2008@gmail.com

**Abstract.** As a mathematical tool for describing random processes, Markov Chain has shown great potential in the field of music analysis. Existing musicology research often focuses on the qualitative analysis of works, while there is still much room for exploration in the quantitative description of musical elements. This study explores the application of Markov chains in music analysis, focusing on contrasting the musical representations of two characters—Carmen and Micaëla—in Georges Bizet's opera Carmen. Using Musical Instrument Digital Interface (MIDI) files, musical elements were encoded into transition matrices for pitch and harmony, capturing the probabilistic relationships between notes and chords. Stationary distributions and entropy were computed to reveal long-term probabilistic behavior and musical complexity, respectively. Results demonstrate that Carmen's music exhibits higher entropy and unpredictable stationary distributions, aligning with her volatile and defiant nature. In contrast, Micaëla's more structured musical transitions reflect her conventional and stable character. By bridging mathematical modeling with artistic interpretation, this study offers a novel quantitative approach to character analysis in opera and lays the groundwork for future interdisciplinary explorations in musicology.

**Keywords:** Markov chains, Music analysis, Bizet's Carmen, Stationary distribution, Entropy.

## 1. Introduction

Mathematical methods have become an invaluable tool in the study of music, offering a structured framework to quantify and analyze complex pieces of all genres. Mathematical tools such as Markov chains, Fourier analysis, and network theory provide objective metrics to uncover musical patterns and structures that might otherwise be ignored. These methods enable researchers to rigorously study the relationships between notes, chords, and rhythms, shedding light on the underlying principles that govern music composition and performance.

Markov Chain is a model that captures transitions between states based on probabilistic rules. In the context of music, Markov chains can describe transitions between notes or chords, offering a systematic way to study the musical properties of all pieces. Researchers have utilized the Markov chain to analyze performers' interpretation of musical elements such as tempo [1]. Based on the Markov chain, Kayser has proposed models to generate music, which exhibits high applicability in real performance [2]. However, there are not many studies relating the Markov chain to pure music analysis of pieces.

Thus, taking Bizet's opera Carmen as an example, this study employs the Markov chain to build a mathematical model of the music and explores its potential to reach more profound conclusions about the opera. Georges Bizet's opera Carmen is a four-act masterpiece, which contrasts the characters of Carmen and Micaëla, illustrating Carmen as "flirtatious" and "volatile," while Micaëla is depicted as "angelic" and "chaste" [3]. By employing the Markov chain, this study aims to apply mathematical lens to contrast the music part and further the characteristics of Carmen and Micaëla. More significantly, it aims at bridging mathematical modeling and artistic music interpretation. Then, it calculates the stationary distribution and entropy of the matrices, gaining insights into the complexity and style of the two characters' parts. Finally, this study attempts to interpret the difference between the two characters' personalities through mathematical results.

# 2. Methodology

## 2.1. Data Description

The dataset for this study includes Carmen's Habanera and Seguidilla and Duet, as well as Micaëla's C'est des contrebandiers and Parle-moi de ma mère. This study chooses these sections as data because all four are rather solo or duet, which means that these music clips best exhibit the characters' features and personalities. For the first three pieces, Musical Instrument Digital Interface (MIDI) files are found online; for the last one, the sheet music is converted to a MIDI file via Optical Music Recognition (OMR), for computational analysis. MIDI files store audio information of notes, which, in this study, will be further extracted and analyzed computationally.

## 2.2. Mathematical Method

To analyze and contrast the lines of Carmen and Micaëla in Bizet's Carmen, this study employs a Markov Chain model, a mathematical tool capable of analyzing several music dimensions (notes, harmony, structure) and identifying probabilistic patterns that reveal the underlying sequential dependencies within a composition or musical corpus. A Markov Chain is a sequence of states in which the probability of transitioning from one state to another depends only on the current state. Mathematically, a Markov Chain can be represented by a transition matrix, where each element $p_{ij}$ represents the probability of transitioning from state i to state j. Thus, the study has [4]:

$$P(X_{n+1} = j \mid X_n = i) = p_{ij}. \tag{1}$$

Then, the study construct four transition matrices, respectively one pitch matrix and one chord matrix for Carmen and Micaëla, namely P1, C1, P2, and C2. The pitch matrix describes the probabilities of transitioning among notes (in the order of C, C#, D, D#, E, F, F#, G, G#, A, A#, B), as shown below:

$$\begin{bmatrix} p_{CC} & \cdots & p_{CB} \\ \vdots & \ddots & \vdots \\ p_{BC} & \cdots & p_{BB} \end{bmatrix}. \tag{2}$$

The harmony matrix describes the probabilities of transitioning among chords (in the order of chord 1, 2, 3, 4, 5, 6, 7), as shown below:

$$\begin{bmatrix} p_{11} & \cdots & p_{17} \\ \vdots & \ddots & \vdots \\ p_{71} & \cdots & p_{77} \end{bmatrix} \tag{3}$$

Note that we use Arabic instead of Roman numerals to avoid the capitalization difference between major and minor keys.

With the matrices, the study calculates the stationary distribution and the entropy for further insights. In the context of Markov Chains, the stationary distribution provides a long-term probabilistic description of the states, offering information on the structural and panoramic differences between Micaëla's and Carmen's lines. More specifically, stationary distribution

eigenvector of the transition matrix. In mathematical language, for a transition matrix P, π = ( π i: i∈I ) is the stationary distribution if:

$$p_{ij}^{(n)} \rightarrow \pi_i \; as \; n \rightarrow \infty \; for \; all \; i \in I \tag{4}$$

Where the state space I is finite [4].

To obtain the stationary distribution of the matrices, the following formula is employed:

$$\pi = P\pi. \tag{5}$$

Then, the entropy will be calculated with the stationary distribution. In information theory, entropy measures the uncertainty of a system. Which, in the context of a Markov Chain, quantifies the diversity of transitions between states. Thus, this study, offers insights into the complexity of transitions among notes and chords, assessing the overall unpredictability and complexity of the two characters' musical parts. Let b be any positive real number and $\pi$ i be the stationary distribution, the entropy of the Markov Chain is defined as [5]:

$$H = -\sum_{i=1}^{n} \pi_i \sum_{j=1}^{n} p_{ij} \log_b(p_{ij}) \tag{6}$$

This study will use 2 as the base to calculate the entropy.

## 3. Results

The matrices computed are shown below, respectively describing transition probabilities between notes and chords of Carmen's and Micaëla's parts.

$$\boldsymbol{P_1} = \begin{bmatrix} 0.11 & 0.01 & 0.10 & 0.39 & 0.07 & 0.04 & 0.00 & 0.02 & 0.01 & 0.01 & 0.18 & 0.05 \\ 0.08 & 0.50 & 0.05 & 0.03 & 0.03 & 0.05 & 0.03 & 0.03 & 0.00 & 0.03 & 0.11 & 0.08 \\ 0.11 & 0.00 & 0.06 & 0.10 & 0.02 & 0.07 & 0.06 & 0.26 & 0.16 & 0.00 & 0.11 & 0.04 \\ 0.01 & 0.01 & 0.08 & 0.08 & 0.00 & 0.07 & 0.26 & 0.26 & 0.03 & 0.00 & 0.17 & 0.02 \\ 0.02 & 0.04 & 0.02 & 0.05 & 0.07 & 0.11 & 0.00 & 0.53 & 0.07 & 0.04 & 0.05 & 0.00 \\ 0.03 & 0.00 & 0.03 & 0.24 & 0.04 & 0.10 & 0.06 & 0.20 & 0.09 & 0.04 & 0.11 & 0.04 \\ 0.50 & 0.00 & 0.02 & 0.03 & 0.02 & 0.02 & 0.06 & 0.03 & 0.10 & 0.04 & 0.10 & 0.08 \\ 0.06 & 0.00 & 0.08 & 0.08 & 0.01 & 0.06 & 0.01 & 0.17 & 0.08 & 0.01 & 0.33 & 0.09 \\ 0.06 & 0.01 & 0.11 & 0.03 & 0.01 & 0.01 & 0.01 & 0.05 & 0.15 & 0.21 & 0.30 & 0.04 \\ 0.16 & 0.00 & 0.01 & 0.08 & 0.04 & 0.01 & 0.00 & 0.03 & 0.52 & 0.00 & 0.08 & 0.07 \\ 0.12 & 0.01 & 0.19 & 0.23 & 0.01 & 0.06 & 0.00 & 0.07 & 0.10 & 0.02 & 0.16 & 0.03 \\ 0.04 & 0.05 & 0.36 & 0.17 & 0.13 & 0.04 & 0.05 & 0.05 & 0.01 & 0.02 & 0.03 & 0.04 \end{bmatrix}$$

$$\boldsymbol{P_2} = \begin{bmatrix} 0.56 & 0.00 & 0.06 & 0.00 & 0.00 & 0.00 & 0.06 & 0.00 & 0.00 & 0.00 & 0.00 & 0.33 \\ 0.14 & 0.08 & 0.04 & 0.00 & 0.00 & 0.02 & 0.04 & 0.40 & 0.00 & 0.20 & 0.00 & 0.08 \\ 0.02 & 0.04 & 0.07 & 0.00 & 0.14 & 0.06 & 0.07 & 0.02 & 0.00 & 0.37 & 0.09 & 0.12 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.04 & 0.28 & 0.00 & 0.00 & 0.07 & 0.19 & 0.15 & 0.00 & 0.20 & 0.00 & 0.07 \\ 0.00 & 0.00 & 0.32 & 0.00 & 0.27 & 0.03 & 0.00 & 0.13 & 0.00 & 0.25 & 0.00 & 0.00 \\ 0.00 & 0.03 & 0.25 & 0.00 & 0.11 & 0.00 & 0.01 & 0.09 & 0.00 & 0.38 & 0.00 & 0.14 \\ 0.02 & 0.04 & 0.05 & 0.00 & 0.28 & 0.12 & 0.09 & 0.00 & 0.00 & 0.09 & 0.07 & 0.24 \\ 0.00 & 0.00 & 0.25 & 0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.00 & 0.25 & 0.00 & 0.00 \\ 0.00 & 0.01 & 0.39 & 0.00 & 0.02 & 0.06 & 0.30 & 0.06 & 0.02 & 0.13 & 0.00 & 0.02 \\ 0.00 & 0.08 & 0.33 & 0.00 & 0.13 & 0.00 & 0.04 & 0.25 & 0.00 & 0.17 & 0.00 & 0.00 \\ 0.00 & 0.04 & 0.19 & 0.00 & 0.14 & 0.00 & 0.05 & 0.44 & 0.01 & 0.04 & 0.05 & 0.03 \end{bmatrix}$$

$$\mathbf{C_1} = \begin{bmatrix} 0.02 & 0.32 & 0.00 & 0.06 & 0.53 & 0.02 & 0.05 \\ 0.29 & 0.00 & 0.00 & 0.29 & 0.38 & 0.05 & 0.00 \\ 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.11 & 0.00 & 0.00 & 0.00 & 0.59 & 0.30 & 0.00 \\ 0.68 & 0.00 & 0.00 & 0.23 & 0.00 & 0.06 & 0.03 \\ 0.47 & 0.07 & 0.07 & 0.07 & 0.27 & 0.00 & 0.07 \\ 0.17 & 0.00 & 0.00 & 0.33 & 0.17 & 0.33 & 0.00 \end{bmatrix} \mathbf{C_2} = \begin{bmatrix} 0.14 & 0.10 & 0.14 & 0.19 & 0.38 & 0.05 & 0.00 \\ 0.33 & 0.00 & 0.33 & 0.00 & 0.33 & 0.00 & 0.00 \\ 0.33 & 0.00 & 0.00 & 0.00 & 0.33 & 0.33 & 0.00 \\ 0.00 & 0.25 & 0.25 & 0.00 & 0.50 & 0.00 & 0.00 \\ 0.93 & 0.00 & 0.00 & 0.00 & 0.00 & 0.07 & 0.00 \\ 0.25 & 0.00 & 0.25 & 0.00 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix} \tag{7}$$

The stationary distributions of $\boldsymbol{P_1}$, $\boldsymbol{P_2}$, $\boldsymbol{C_1}$, $\boldsymbol{C_2}$ are respectively:

$$\begin{aligned} \pi_{P1} &= [0.10,0.02,0.10,0.14,0.03,0.06,0.06,0.14,0.10,0.04,0.18,0.05] \\ \pi_{P2} &= [0.02,0.03,0.21,0.00,0.11,0.05,0.12,0.12,0.01,0.21,0.03,0.09] \\ \pi_{C1} &= [0.32,0.11,0.01,0.14,0.32,0.08,0.03] \\ \pi_{C2} &= [0.40,0.06,0.11,0.08,0.28,0.08,0.00] \end{aligned} \tag{8}$$

The entropy of $\boldsymbol{P_1}$, $\boldsymbol{P_2}$, $\boldsymbol{H_1}$, $\boldsymbol{H_2}$ are respectively: $H_{P1}$ = 33.00; $H_{P2}$ = 24.98; $H_{C1}$ 10.02; $H_{C2}$ = 8.86.

In the results, relatively high entropy and unpredictable stationary distribution indicate more complex musicality and potentially more volatile personality; in contrast, relatively low entropy and predictable stationary distribution tell more conventional musical structure and melodies, and likely a more obedient character.

The stationary distribution πP1 reveals relatively balanced probabilities across most notes, with only one relatively dominant peak at A#/Bb (0.18). In contrast, πP2 exhibits the significant dominance of D (0.21) and G (0.21), with much lower probabilities at other states. Then, the entropy provides a higher-level summary of the overall behavior of the two matrices. With HP1 (33.00) higher than HP2 (24.98), it suggests the overall high variety of notes in Carmen's part that reflects her complex and volatile personality, while indicating a more predictable melodic structure of Micaëla that aligns with her character as a "conventional European woman".

In addition, πC1 and πC2 both project an emphasis on chord 1 (respectively 0.32 and 0.40) and chord 5 (respectively 0.32 and 0.28). This is consistent with the rule of the authentic cadence in Western composition, validating the accuracy of the result. Notably, the long-term probability of arriving at chord 1 in Carmen's part (0.32) is smaller than that in Micaëla's (0.40), meaning that, in Micaëla's part, the music has a stronger tendency to return to the tonic. This also tells Micaëla's characteristics as more obedient and traditional. Furthermore, the results of the entropy – HC1 (10.02) higher than HC2 (8.86) – indicate the unpredictability of Carmen's part in terms of chord progression.

## 4. Discussion

The integration of Markov Chains and musical analysis bridges quantitative and qualitative methodologies. By interpreting notes and chords as states in a probabilistic system, this study creates mathematical note-space and chord space. The stationary distribution measures the music's tonal and harmonic emphasis, offering a quantitative understanding of musical motif and recurrence. In addition, the entropy quantitatively assesses the complexity of music, also translating abstract musical language into a measurable system. Furthermore, the result's alignment with Western musical norms (e.g. authentic cadences) validates the computational and mathematical approach toward music analysis.

The contrast in stationary distributions and entropy between the two characters underscores the use of musical semiotics as a narrative tool in the opera. Carmen's relatively unpredictable transitions, both in pitch and harmony, allude to her defiant and free-spirited personality. Conversely, Micaëla's relatively structured musical transitions align with her role as a representation of conventional values and stability. These conclusions from mathematical results are consistent with mature humanities studies on the opera Carmen [3].

## 5. Conclusion

By employing Markov chains, this study adopts a mathematical framework to dissect and compare the contrasting musical languages of Carmen and Micaëla, two important characters in Bizet's Carmen. The approach leverages a fundamental property of music: it progresses by ascending or descending intervals based on the previous note, which highly resembles the probabilistic properties of Markov chains. By identifying notes and chords as discrete states in the chain, the study builds a mathematical model for music analysis.

The process begins with the conversion of musical elements into a computationally analyzable format through MIDI files. Then, the study proceeds to the creation of transition matrices that encode the probabilities of moving between states – notes and chords. These matrices serve as the foundation for further mathematical analysis, including the computation of stationary distributions and entropy.

The stationary distribution provides the long-term probabilistic behavior of states, revealing the relative importance of certain notes or chords in the music. This offers insights into the structure and focal points within each character's part. The entropy, in addition, quantifies the unpredictability of the transition matrices, serving as a metric for assessing the complexity of the music.

The quantitative results showcase the characters' traits and provide sharp contrast. For Carmen, the high entropy and unpredictable stationary distribution align with her free-spirited, volatile, and defiant nature. Conversely, Micaëla's music exhibits lower entropy and a more predictable stationary distribution, reflecting her conventional demeanor.

This study introduces an innovative quantitative approach to character analysis in opera. The approach is potentially also applicable in other types of music analysis such as cross-genre comparison. However, this study only uses the first-order Markov chain, which does not account for longer-term dependencies. In the future, second-order or higher-order Markov chains could be incorporated to capture more complex dependencies between musical states.

## References

[1] McDonald, D. J., McBride, M., Gu, Y., & Raphael, C. Markov-switching state space models for uncovering musical interpretation. The Annals of Applied Statistics, 2021, 15(3): 1147-1170.

[2] Kayser, M. Generative models of music. 2013. Retrieved from http://cs229.stanford.edu/proj2013/Kayser-GenerativeModelsOfMusic.pdf

[3] McClary, S. Georges Bizet: Carmen. 1992. Retrieved from https://doi.org/10.1017/cbo9781139166416

[4] Norris, J. R. Markov chains. 1998. Cambridge University Press.

[5] LeBlanc, P. Information theory: Entropy, Markov chains, and Huffman coding. n.d.