PerFormer: An AI-Driven Approach to Melody Generation in Microtonal Persian Music

Farzad Hosseinabadi
farzad.hosseinabadi@hud.ac.uk
University of Huddersfield
Huddersfield, UK

Ian Gibson
i.s.gibson@hud.ac.uk
University of Huddersfield
Huddersfield, UK

Christopher P. Dewey
c.dewey@hud.ac.uk
University of Huddersfield
Huddersfield, UK

Abstract

This paper presents PerFormer, a Transformer-based framework for generating monophonic microtonal melodies in Persian classical music. While most symbolic music generation systems are designed for the twelve-tone equal temperament (12-TET) system, many musical traditions employ microtonal pitch structures that fall outside this framework. Persian classical music is a prominent example, characterised by monophonic textures and distinctive intervallic systems that are not adequately supported by existing generative models.

To address this gap, we introduce a culturally informed event-based symbolic representation that encodes microtonal pitch categories, including Persian accidentals such as *sori* (quarter-tone sharp) and *koron* (quarter-tone flat), together with rhythmic duration and intra-measure positional information. The model is trained on a curated dataset of Persian melodies, augmented through controlled microtonal transposition, and represented as position–pitch–duration token sequences.

The system is evaluated using both quantitative metrics and a perceptual listening study in comparison with a first-order Markov baseline. Quantitative results show that the model preserves modal pitch inventories, rhythmic duration characteristics, and metric position distributions, while perceptual evaluation indicates substantial improvements in musical coherence, stylistic similarity, pitch correctness, and overall quality.

These findings indicate that Transformer architectures can be effectively adapted to microtonal musical traditions when supported by appropriate representational design. The proposed framework provides a foundation for future work in computational tools for improvisation, composition, and pedagogy within culturally diverse musical systems.

Keywords

Melody Generation, Persian Classical Music, Microtonality, Symbolic Music, Artificial Intelligence in Music

1 Introduction

Deep learning has become a central methodology in computational music generation, with recent advances driven largely by sequence modelling architectures. Early approaches based on recurrent neural networks (RNNs), including Long Short-Term Memory (LSTM) models demonstrated promising results in modelling temporal musical structure [1], [2], [3]. However, RNN-based systems are inherently limited in capturing long-range dependencies due to vanishing gradients and constrained internal memory. These limitations are particularly problematic in music, where large-scale structure, phrase development, and rhythmic organisation often extend over long temporal spans.

Transformer architectures [4] address these challenges through self-attention mechanisms that allow direct modelling of relationships between distant elements in a sequence. By removing the sequential bottleneck of recurrence, Transformers can learn hierarchical musical structures more effectively and have achieved state-of-the-art performance in a variety of music generation tasks. Prior work has also shown that Transformer-based systems can produce music with improved coherence, stylistic consistency, and expressive variation, as measured by both quantitative metrics and human evaluation [5].

Despite this progress, most research in neural music generation remains heavily focused on Western tonal traditions and the twelve-tone equal temperament (12-TET) system. Microtonal musical systems, which employ pitch structures beyond standard semitone divisions, remain significantly underrepresented. Persian classical music is a complex microtonal tradition characterised by modal systems (dastgāh), nuanced pitch inflections, and distinctive rhythmic and melodic conventions. These features present unique challenges for computational modelling, including the representation of microtonal pitch categories, phrase structure, and metric organisation.

This paper introduces PerFormer, a Transformer-based framework designed specifically for melody generation in microtonal Persian music. The proposed system incorporates a custom event-based representation that encodes microtonal pitch categories, rhythmic duration, and intra-measure positional structure. In addition to the modelling framework, the work presents a curated and microtonally annotated symbolic dataset of Persian monophonic melodies, addressing the scarcity of computational resources for non-Western musical traditions. By integrating culturally informed encoding strategies with a modern sequence modelling architecture and a purpose-built dataset, PerFormer aims to bridge the gap between contemporary AI music generation techniques and underrepresented musical systems.

The contributions of this work are threefold:

- A curated and microtonally annotated symbolic dataset of Persian melodies that preserves pitch and rhythmic structure.
- A novel event-based representation tailored to Persian microtonal music.
- A Transformer-based generative model capable of learning stylistic pitch and rhythmic patterns.

In addition to its technical contributions, the proposed framework suggests potential applications in interactive music systems. By generating stylistically coherent microtonal melodies, PerFormer may support performance-oriented and pedagogical scenarios, including real-time improvisational interaction between human performers and computational agents. Although the current system operates offline, this work provides a basis for future exploration of interactive generative models.

2 Background

This section reviews Transformer architectures and prior work in symbolic music generation. Transformers employ self-attention to model dependencies between tokens in a sequence [4]. The attention mechanism is defined as:

$$A = \mathrm{softmax}\left(\frac{K^{T}Q}{\sqrt{d}}\right) \in R^{m\times m}$$

where Q and K denote query and key matrices, d is the embedding dimensionality, and A represents attention weights describing relationships between all tokens in the sequence. This mechanism enables efficient modelling of long-range musical structures.

Transformer-based systems such as Music Transformer, Pop Music Transformer, and Compound Word Transformer explore event-based representations for expressive music generation [5], [6], [7]. Additional work investigates piano inpainting for interactive composition [8], graph-based modelling of melodic structure [9], motif-conditioned generation [10], and hierarchical melody synthesis [11]. Jazz Transformer extends these ideas to homophonic generation using chord-level representations [12]. Most of these approaches remain grounded in 12-TET frameworks.

Research on microtonal music generation remains relatively limited, with existing work concentrating on a small number of traditions, particularly Turkish music [3], [13]. Turkish music employs a 53-tone division of the octave, a standard Turkish tuning framework. Persian music, by contrast, does not adopt this system, and its modal organisation, based on dastgāh and gūsheh, diverges substantially from the Turkish makam structure. Therefore, Persian music's modal organisation and pitch categories require dedicated modelling strategies. This gap motivates the development of approaches specifically tailored to Persian melodic practice.

3 Methodology

This study addresses the problem of generating microtonal melodies in Persian classical music by combining a culturally informed symbolic representation with a Transformer-based sequence modelling framework.

The central challenge lies in capturing non-12-TET pitch organisation, modal structure, and metric behaviour within a unified computational model. To overcome limitations of conventional symbolic representations, the proposed method encodes musical information as structured event sequences that jointly represent pitch, rhythmic duration, and intra-measure position. This design allows the model to learn relationships between microtonal pitch categories and rhythmic placement, which are essential for reproducing stylistically coherent melodic behaviour. By incorporating a representation aligned with theoretical principles of Persian music, the system enables learning of modal pitch inventories and intervallic patterns that are not directly expressible in standard Western encodings.

Figure 1 illustrates the overall pipeline, including data representation, model training, and melody generation stages.

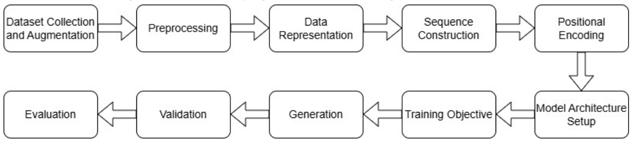

Figure 1- Flow diagram of the proposed modelling pipeline

The generative model follows an autoregressive sequence prediction paradigm in which each musical event is predicted conditioned on all preceding events. Unlike local or Markovian approaches, the Transformer architecture employs self-attention mechanisms that provide global contextual access, enabling the modelling of long-range dependencies such as phrase development, contour consistency, and rhythmic organisation. This is particularly important in Persian music, where melodic structure and modal identity unfold over extended temporal spans.

To support tonal consistency, a tonic reference sequence is incorporated, functioning analogously to a drone and guiding the model in learning relative pitch relationships. This design is inspired by the open-string drone practice (Wākhān) in Persian instruments, where a constant pitch is sustained alongside the melody.

Model performance is evaluated using a combination of quantitative metrics and perceptual listening tests. Quantitative evaluation measures the preservation of modal pitch constraints, rhythmic distributions, and metric position patterns, while perceptual evaluation assesses musical coherence, stylistic similarity, pitch correctness, and overall quality.

4 Dataset

Persian classical music is structured around a set of principal modal scales. Among these modal systems, Chāhārgāh in the key of C is selected as the reference modal framework for analysis and modelling. Chāhārgāh is characterised by a distinctive intervallic structure and expressive melodic behaviour (see Figure 2).

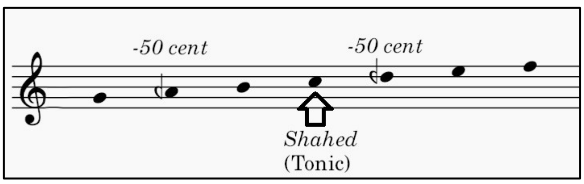

Figure 2- Pitch degrees of the Chāhārgāh modal scale in the key of C

Concentrating on a single modal framework allows the model to learn stylistically coherent pitch and rhythmic patterns within a well-defined musical context.

The dataset consists of 33 metric monophonic Persian melodies curated from radif collections, which document canonical Persian melodic repertoire organised by modal scale systems. These are complemented by selected folk and composed pieces written during the past century, expanding the stylistic range of the dataset [14], [15], [16], [17], [18]. The dataset focuses on metric Persian dance forms such as Reng and Chahārmezrāb, which commonly employ a 6/8 meter. Although these forms differ in tempo, they share comparable stylistic characteristics and metric structures.

5 Data Representation and Model

This section describes the symbolic representation used to encode the dataset, followed by the model architecture, training procedure, and generation process.

5.1 Symbolic Representation

Because Persian musical intervals do not strictly conform to the twelve-tone equal temperament (12-TET) system, an alternative equal-tempered framework is required to discretise pitch space for computational representation. In this study, the 24-tone equal temperament (24-TET) system proposed by Vaziri is adopted [19], [20]. This system divides the octave into 24 equal parts, providing a practical approximation of Persian microtonal intervals.

To construct the symbolic vocabulary for model training, a REMI-like representation [6], [21], [22] adapted for quarter-tone intervals is employed. Three token categories are defined:

1) **Position tokens:** Each 6/8 measure is divided into twelve sixteenth-note positions. Since different metric positions exhibit distinct rhythmic and expressive characteristics in Persian music, twelve position tokens are introduced to encode within-measure location.
2) **Pitch tokens:** The pitch range is selected to reflect the practical register of Persian instruments, spanning from C2 to B_sori5. Under the 24-TET system, each octave contains 24 pitch degrees. Including all pitch classes within this range, together with a dedicated rest symbol, results in a total of 97 pitch-related tokens.
3) **Rhythmic (duration) tokens:** The minimum rhythmic unit is defined as a sixteenth note. By incorporating dotted values, eight discrete rhythmic categories are used to represent note durations. Durations exceeding a dotted half note are segmented into a dotted half note plus the remaining duration, which is represented using the available shorter rhythmic units.

Combining these categories with pad, start, and end tokens yields a total vocabulary of 118 tokens for training. Each melody is encoded as a sequence of Position-Pitch-Duration events, forming an event-based structure suitable for Transformer-based learning.

5.2 Model Architecture and Training

This project employs an encoder–decoder Transformer architecture trained on triplet token sequences to generate Persian microtonal melodies. To increase dataset diversity and improve model generalisation, each melody is augmented through systematic quarter-tone transposition, which is analogous to semitone transposition in [23], [24]. Each piece is transposed by up to 14 quarter-tones upward and 9 quarter-tones downward within the valid pitch range of the representation, covering all supported tonalities. After augmentation, the corpus expands to 590 training sequences. The dataset was split into training (80%), validation (10%), and test (10%) subsets.

To encode the modal pivot (tonic), a tonic reference sequence is included as part of the encoder input, providing a conditioning signal for relative pitch learning. The decoder then generates the melody autoregressively while attending to this encoded tonal context through cross-attention.

The resulting sequences are segmented into fixed-length windows of 256 tokens. Given the Position–Pitch–Duration representation, this corresponds to approximately 85 musical events per sequence. The model employs an embedding dimension of 256, consists of 6 layers, and operates with a vocabulary size of 118. Training is performed with a batch size of 32 and a dropout rate of 0.1. The model is optimised using the Adam optimiser with a learning rate of 1×10^{-4} for 100 epochs on an NVIDIA Tesla T4 GPU. Figure 3 presents the training and validation accuracy and loss curves, indicating stable convergence.

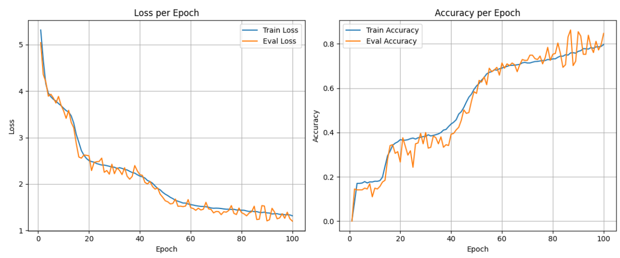

Figure 3- Training accuracy and loss curves

5.3 Generation Procedure

During generation, the model produces microtonal melodies through autoregressive decoding. The resulting token sequences are converted into symbolic scores using the MusicXML format, which supports microtonal accidentals and related metadata. Pitch tokens are sampled using top-k sampling ($k = 20$, temperature = 1.0), while duration tokens use a higher temperature (1.2) to encourage rhythmic diversity. Figure 4 presents an example of a generated score.

6 Evaluation Methods

6.1 Quantitative Evaluation

Several domain-specific musical metrics were employed to quantitatively evaluate the generated melodies, including pitch conformity, pitch consonance, melody–chord tonal distance, and beat and downbeat accuracy [6], [11], [12], [23], [25]. From these, three adapted metrics were selected to analyse key characteristics of the generated outputs. Evaluation was conducted on a set of 20 generated melodies.

Metric position distribution: Metric forms such as Reng and Chahārmezrāb exhibit characteristic patterns of rhythmic emphasis across specific metric positions. To evaluate this aspect, the distribution of note onsets across the twelve sixteenth-note positions within each 6/8 measure was analysed for both PerFormer and a first-order Markov baseline initialised with a short seed sequence derived from the Chahārgāh mode in C to provide tonal context and ensure a fair comparison (Table 1).

The Markov baseline produces an approximately uniform distribution, indicating that it fails to capture metrical hierarchy or rhythmic emphasis due to its reliance on local duration transitions without measure-level awareness. In contrast, PerFormer shows strong concentration on a small set of metrically salient positions, particularly at the beginning of the measure and other regularly spaced anchor points. These peaks align closely with the dataset, indicating successful modelling of metric structure.

Table 1- Comparison of normalised position usage across PerFormer, the baseline model, and the dataset

Position usage comparison (normalized)			
Pos	Markov (Baseline)	PerFormer	Dataset
0	9.29%	21.65%	20.80%
1	7.18%	1.82%	0.41%
2	10.59%	4.71%	4.85%
3	7.00%	3.59%	8.26%
4	9.24%	20.41%	16.85%
5	6.82%	1.35%	0.22%
6	9.94%	12.82%	19.42%
7	6.94%	1.41%	0.62%
8	9.24%	20.65%	17.92%
9	6.53%	1.94%	1.40%
10	10.00%	7.76%	8.38%
11	7.24%	1.88%	1.15%

Allowable pitch accuracy: Since Persian classical practice does not employ Western-style modulation, each dastgāh (e.g., Chāhārgāh) is characterised by a relatively stable pitch inventory. This enables a constraint-based metric measuring the proportion of generated pitch tokens within a predefined allowable set for the target tonic.

Table 2 presents results for PerFormer and a first-order Markov baseline. Using a dataset augmented across all 24 tonal centres, the Markov model achieves a mean allowable-pitch accuracy of 34.71%, indicating frequent violations of the modal inventory. This reflects the limitation of modelling only local pitch transitions without tonal context.

In contrast, PerFormer achieves 90.18% allowable-pitch accuracy. This improvement is partly supported by a tonic reference provided through a continuous drone melody, and by the model's ability to learn relative relationships between pitch degrees across different tonalities. Despite operating over the full pitch vocabulary without explicit constraints, the model effectively captures tonic-dependent modal structure. The lower standard deviation for PerFormer (3.69% vs. 4.46%) further indicates more consistent behaviour across generated samples.

Table 2- Allowable pitch percentages for PerFormer and the baseline model

Summary of allowable pitch		
Metric	Markov (Baseline)	PerFormer
Mean allowable pitch	34.71%	90.18%
Std. allowable pitch	4.46%	3.69%
Min. allowable pitch	25.88%	83.53%
Max. allowable pitch	44.71%	95.29%

Rhythmic duration distribution: In addition to pitch and metric position, the normalised distribution of rhythmic durations in the generated melodies is compared with the dataset in Table 3. The generated outputs exhibit a shift toward longer note values, with quarter notes occurring substantially more frequently than in the dataset. In contrast, shorter durations, particularly sixteenth and eighth notes, appear less often than expected. Several extended rhythmic values (e.g., dotted quarter and half notes) are present, but they remain relatively infrequent. The overall similarity between the generated and dataset duration distributions was computed using cosine similarity yielding a value of 86%, indicating that key aspects of the rhythmic profile are preserved, alongside a tendency toward rhythmic simplification and lengthening of note values.

Table 3- Average normalised rhythmic duration distribution between PerFormer and the dataset

Average duration distribution		
Duration name	PerFormer	Dataset
16th note	0.0729	0.1491
8th note	0.3971	0.5322
dotted 8th note	0.0618	0.0879
quarter note	0.4312	0.1652
quarter note + 16th	0.0029	0.0000
dotted quarter note	0.0118	0.0593
dotted quarter note + 16th	0.0088	0.0016
half note	0.0135	0.0074

These results collectively demonstrate that modelling long-range dependencies and tonal relationships is essential for generating coherent and stylistically accurate microtonal music.

6.2 Qualitative Evaluation

Visual evaluation: Figure 4 presents a visual and musical inspection of one representative example. The generated melody preserves the modal pitch inventory and characteristic intervallic patterns consistent with the Chāhārgāh style. While the overall contour is generally plausible, the phrasal structure is less consistently coherent.

Figure 4- Example of a generated melody

In this example, 92.94% of the notes fall within the predefined allowable pitch set, while the remaining out-of-set notes, highlighted with red rectangles, represent deviations from the modal constraints. The excerpt also demonstrates a varied use of rhythmic values and pitch movement, indicating that the model produces musically diverse melodic material.

Subjective evaluation: A perceptual evaluation was conducted using 10 randomly ordered 15-second audio excerpts, comprising 5 generated by PerFormer and 5 by a first-order Markov baseline. Two musicians with experience in Persian classical music participated in the study. Participants rated each sample on a 5-point Likert scale across four criteria: musical coherence (melodic structure), stylistic similarity to Persian classical music, pitch correctness (microtonality), and overall quality.

Table 4 shows that PerFormer consistently outperforms the Markov baseline across all perceptual criteria, with the largest improvement observed in stylistic similarity and pitch correctness. Musical coherence improved from 2.2 to 3.9, indicating stronger structural organisation. Stylistic similarity increased from 1.4 to 4.4, demonstrating effective modelling of Persian musical characteristics. Pitch correctness also improved significantly (1.4 to 4.1), reflecting the model's ability to capture microtonal relationships. Overall quality increased from 1.5 to 3.9, confirming that the generated outputs are perceptually more convincing.

Table 4 - Results of subjective evaluation comparing PerFormer and the Markov baseline

Metric	Markov (Baseline)	PerFormer
Musical coherence (melodic structure)	2.2	3.9
Stylistic similarity (Persian classical music)	1.4	4.4
Pitch correctness (microtonality)	1.4	4.1
Overall quality	1.5	3.9

These results suggest that modelling long-range dependencies and relative tonal relationships is essential for generating coherent and stylistically accurate microtonal music.

7 Conclusion

This study introduced PerFormer, a Transformer-based system for generating monophonic microtonal melodies in Persian classical music. By combining a culturally informed symbolic representation with a state-of-the-art neural sequence modelling architecture, the framework addresses limitations in existing music generation research, which has largely focused on Western twelve-tone equal temperament systems. Quantitative and qualitative evaluations indicate that the model preserves key modal and rhythmic characteristics of the Chāhārgāh repertoire, suggesting that Transformer architectures can be effectively adapted to microtonal musical traditions when supported by appropriate representational design.

8 Practical Implications and Future Work

Although the current system operates offline, it offers a basis for future real-time interactive applications, particularly in musical improvisation. The model could be integrated into environments where a performer presents a melodic phrase and receives a responsive continuation. This interaction may support performance practices and serve as a pedagogical tool for understanding modal structures and improvisational strategies. The model may also function as a compositional aid by proposing stylistically consistent melodic responses within microtonal frameworks.

Future work will extend the framework to additional dastgāh systems, incorporate higher-level phrase modelling, and explore interactive implementations. Expanding the dataset and conducting larger-scale listener-based evaluations will further clarify the musical validity and creative potential of microtonal generative systems.

9 Limitations

The dataset is relatively small and limited to a single dastgāh and metric style, which may restrict generalisability despite augmentation across tonal centres. In addition, the current system operates offline and does not yet support real-time interaction.

Ethical Standards

This research did not involve human participants, human data, or animal subjects. The dataset is derived from published and publicly available sources and is used for non-commercial research purposes only; the original material is not redistributed. No personal or sensitive data were collected or processed.

The authors declare no financial or non-financial conflicts of interest.

This research did not receive any specific external funding.

References

[1] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: a Steerable Model for Bach Chorales Generation," 2017. [Online].

[2] C. Benetatos, J. Vanderstel, and Z. Duan, "BachDuet: A Deep Learning System for Human-Machine Counterpoint

Improvisation," 2020. [Online]. Available: http://www2.ece.rochester.edu/projects/air/

[3] A. Kaşif and S. Sevgen, "Classical Turkish Music Composition with LSTM Self-Attention," *Journal of Innovative Science and Engineering (JISE)*, Feb. 2024, doi: 10.38088/jise.1406162.

[4] A. Vaswani *et al.*, "Attention Is All You Need," 2017.

[5] C.-Z. A. Huang *et al.*, "Music Transformer," Dec. 2018, [Online]. Available: http://arxiv.org/abs/1809.04281

[6] Y. S. Huang and Y. H. Yang, "Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2020, pp. 1180–1188. doi: 10.1145/3394171.3413671.

[7] W.-Y. Hsiao, J.-Y. Liu, Y.-C. Yeh, and Y.-H. Yang, "Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs," Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.02402

[8] G. Hadjeres and L. Crestel, "The Piano Inpainting Application," Jul. 2021, [Online]. Available: http://arxiv.org/abs/2107.05944

[9] Y. Zou, P. Zou, Y. Zhao, K. Zhang, R. Zhang, and X. Wang, "MELONS: GENERATING MELODY WITH LONG-TERM STRUCTURE USING TRANSFORMERS AND STRUCTURE GRAPH," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 6542–6546. doi: 10.1109/ICASSP43922.2022.9747802.

[10] Y.-J. Shih, S.-L. Wu, F. Zalkow, M. Müller, and Y.-H. Yang, "Theme Transformer: Symbolic Music Generation with Theme-Conditioned Transformer," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2111.04093

[11] Y. Lv, J. Luo, B. Ju, and X. Yang, "Small Tunes Transformer: Exploring Macro & Micro-Level Hierarchies for Skeleton-Conditioned Melody Generation," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.08626

[12] S.-L. Wu and Y.-H. Yang, "The Jazz Transformer on the Front Line: Exploring the Shortcomings of AI-composed Music through Quantitative Measures," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.01307

[13] I. H. Parlak, Y. Çebi, C. Işikhan, and D. Birant, "Deep learning for Turkish makam music composition," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 29, no. 7, 2021, doi: 10.3906/ELK-2101-44.

[14] Dariush Talai, *Radif of Mirza Abdollah*. Tehran: Ney, 2023.

[15] Manouchehr Lashgari, *Radif of Iranian Music for Violin, Kamancheh, Flute and Ney*. Chang, 2019.

[16] Hossein Alizâdeh, *The Târ and Setâr Intermediate Course*. Mahoor, 2000.

[17] Jalal Zolfonun, "Setar Playing Tutorial," *Hastan*, vol. 3, 2010.

[18] Arshad Tahmâsbi, *Works of Darvish Khan*. Mahoor, 2013.

[19] Ali-Naqi Vaziri, *Music Theory*, 10th ed. Safi Alishâh, 2023.

[20] Ali-Naqi Vaziri, *The Harmony of the Music of Iran or Quarter-tone Music*. Mahoor, 2016.

[21] S.-L. Wu and Y.-H. Yang, "MuseMorphose: Full-Song and Fine-Grained Piano Music Style Transfer with One Transformer VAE," Dec. 2022, [Online]. Available: http://arxiv.org/abs/2105.04090

[22] B. Yu *et al.*, "Museformer: Transformer with Fine- and Coarse-Grained Attention for Music Generation," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.10349

[23] J. Zhou, H. Zhu, and X. Wang, "CHOIR TRANSFORMER: GENERATING POLYPHONIC MUSIC WITH RELATIVE ATTENTION ON TRANSFORMER," 2023. [Online]. Available: https://github.com/Zjy0401/choir-transformer

[24] L. Ø. Bentsen, R. Simionato, B. Wallace, and M. Krzyzaniak, "Transformer and LSTM Models for Automatic Counterpoint Generation using Raw Audio," 2022.

[25] S. Wu, Y. Wang, X. Li, F. Yu, and M. Sun, "MelodyT5: A Unified Score-to-Score Transformer for Symbolic Music Processing," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.02277