Ethnic Music Style Migration Algorithm Based on Artificial Intelligence

Beini Liu

Abstract

This paper introduces a novel approach that is designed for the purpose of enhancing the migration of Ethnic Music (EM) styles into diverse musical content using Artificial Intelligence (AI) techniques. EM is considered to be rich in cultural significance. It is characterized by unique scales, rhythms, and instruments, so it presents specific challenges in style transfer, mainly in terms of preserving the authenticity and integrity of traditional sounds while, at the same time, trying to introduce new stylistic elements. The proposed Migration Algorithm (MA) employs a combination of techniques such as Short-Time Fourier Transform (STFT) for preprocessing, Feature Extraction (FE) using Mel scale filtering and Mel-Frequency Cepstral Coefficients (MFCC), and feature transformation modified convolutional neural networks (STLCNet for style and CONCNet for content). The output from both networks is combined using the mathematical theory of evidence, and its output spectrogram is reconstructed into an audio signal using the phase gradient heap integration (PGHI) algorithm. An additional 60,000 recordings have been included in the Smithsonian Folkways Collection for testing. The approach succeeded significantly on key metrics, including KL Divergence, Signal-to-Distortion Ratio (SDR), and Signal-to-Noise Ratio (SNR), and the empirical findings proved that type of transfer accuracy, quality of sound, and text integrity were all properly addressed.

Keywords: Ethnic Music, Artificial Intelligence, Short-Time Fourier Transform, Quality of Audio, Signal-to-Noise Ratio, Accuracy

INTRODUCTION

The phrase "Ethnic Music (EM)" describes an infinite number of types of music that have their respective basis in the culture and traditions of specific ethnic groups. Maintaining the community's language, festivals, and historical events represents the social and historical facts of the community. Through the method of musical leaving their homes, EM can be included in deeper musical panoramas, which in response supports intercultural communication and refinement. Musicians perform type migration when they adopt features from one musical style and implement them in a new one. Presenting an atmosphere for innovative expression through music allows musicians to test themselves musically while maintaining their inspiration from an extensive cultural past in their distinctive tunes. Migrating types with the assistance of Artificial Intelligence (AI) techniques enable EM to be even more available and permanent.

Traditional music evolution versions, particularly those related to EM, present unique challenges, such as requiring the inclusion of new musical elements while maintaining the true spirit of the original music. The primary limitation is how well it communicates the key elements of the original score, particularly its unique boundaries, complex tempos, and orchestration. Conventional Western musical notes and concepts used in most style import algorithms frequently ignore the extensive range and complexity that comprise EM, rendering the process even more challenging.

Findings that attempt to adequately represent the core of the first approach and mislead critical cultural factors are prevalent when using modern algorithms, which may not be appropriate to this purpose because they cannot model and reproduce these complex features sufficiently. In addition, the complexity and range of EM histories disappear in the simplified versions that result from currently available methods’ failure to compensate for the non-linear relationships between different musical elements sufficiently. Because of the limitation, the final result is of less high quality, and the music drops some of its social and historic significance owing to the deletion of key elements.

A powered by Artificial Intelligence (AI)-Migration Algorithm (MA) for EM is presented in this paper. The algorithm starts by changing the input data into frequency-domain illustrations utilizing the Short-Time Fourier...
Transform (STFT), which is employed for initial processing input audio signals according to composition and style. The following process extracts features from the input style and audio content using Mel-Frequency Cepstral Coefficients (MFCC) and Mel scale filtering processes. Two CNNs, CONCNet for quality audio input and STLCNet for style, have been implemented to transform these extracted features. Then, a fused audio spectrogram is created by integrating the two networks’ results and applying the Mathematical Theory of Evidence technique. This spectrogram is then transformed into an audio signal using the Phase Gradient Heap Integration (PGHI) algorithm. For the experimental analysis, the study employed the Smithsonian Folkways Collection database, which contains over 60,000 tracks of various ethnic and cultural groups. The proposed model was examined for several key metrics in the analysis, and its effectiveness in style transfer and quality of audio (QoA) was demonstrated. The Signal-to-Distortion Ratio (SDR) hits 20 dB, and the Signal-to-Noise Ratio (SNR) hits 27 dB with a KL deviation level as low as 0.25.

The structure of this study paper is as outlined below: A review of the literature is given in Section 2, the model suggested in Section 3, the investigation of the tests in Section 4, and the work ends in Section 5.

LITERATURE REVIEW

A novel approach based on intelligent neural networks for composing music is described by [1]. Using a reward-based method for sequence probability fixation and feedback systems, the researchers used the performer's long short-term memory (LSTM) to develop musical episodes. In a related research project that examined traditional music trends filtering, investigators discovered that the neural model had challenges with feature selection and that hypothetical music guidelines may have assisted with trend adhesion. These guidelines were subsequently approved through individual tests.

[2] attempted to address these issues by identifying traditional musical styles using Artificial Neural Networks (ANN). In order to accomplish this, they encoded the audio signals into sound patterns and employed algorithms to extract style-specific elements that had proven superior precision in the experimental study.

In order to successfully recommend music, [3] developed a collaborative filtering algorithm-based approach that incorporated into account user features, music styles, and chronological interests. Their strategy showed to be far more successful in improving platform management and user comfort, based on their test study.

In their research, [4] revealed Groove2Groove, an approach for symbolic music transfer that relies on one-shot style transfer and focuses on accompaniment styles in traditional and jazz styles. Additional synthetic data collection enabled experts to investigate their neural network architecture, and results showed better approximation and style transfer performance.

Applying an approach referred to as vector-quantized variational autoencoders, [5-6] introduced a one-shot timbre transfer method for addressing this gap. Their approach employed self-supervised learning for the task of music transfer analysis, and the model achieved superior results in timbre manipulation compared to existing methods.

[7-10] introduced ChordGAN, and the model utilized conditional GAN architecture to transfer harmonic structures across music genres. The method embedded chroma feature extraction and also employed loss functions to transfer style elements accurately and, at the same time, maintain the content fidelity.

PROPOSED MODEL

The proposed Music style MA is presented in Fig. 1. The architecture involves first processing the style audio and content audio input through STFT, after which the process is followed through power spectrum calculation for the style audio and Mel scale filtering for the content audio. The output from the style audio pipeline is then fed to the Log-Mel spectrogram creation, and the Content audio pipeline is fed to the MFCC pipeline. The output from both pipelines is then processed using the modified convolutional network models, such as the STLCNet and CONCNet [11-12], for style and content audio input, respectively. Next, in the process flow, the output from both networks is fused using the mathematical theory of evidence, creating a fused spectrogram.
To conclude the music form synthesis analysis, the framework utilizes the Phase Gradient Heap Integration method using the fused spectrogram data [13-15].

![Proposed Music MA.](image)

**Figure 1.** Proposed Music MA.

**Preprocessing Using STFT for Style and Content Audio**

Implementing the Short-Time Fourier Transform (STFT) to audio data's composition and subject matter is the primary preprocessing process in the music style MA. Here, the audio is transformed from a time field type to a frequency-domain one. This practice uses a window function and a frequency- and time-dependent conversion to the two audio formats. For both the style audio, denoted as $s(n)$, and the content audio, denoted as $c(n)$, a Hanning window is employed to minimize spectral leakage. The mathematical representation of this window is given by EQU (1).

$$w(n) = 0.5 - 0.5 \cos \left( \frac{2\pi n}{N-1} \right)$$  \hspace{1cm} (1)

where $N$ indicates the window length. The choice of window length and the overlap between successive windows differ based on the audio type to optimize their respective spectral analyses. Specifically, the style audio uses 1024 samples with a 50% overlap window length. In contrast, the content audio utilizes a longer window of 2048 samples with a 75% overlap. The STFT for each audio input is computed using the EQU (2).

$$X(m, \omega) = \sum_{n=-\infty}^{\infty} x(n)w(n - mR)e^{-j\omega n}$$  \hspace{1cm} (2)

where $x(n)$ represents the audio signal under analysis, $R$ is the hop size, and $m$ and $\omega$ index the time frames and frequency components, respectively. With a typical sampling rate of 44.1 kHz for both inputs, the frequency resolution $\Delta f$ and the time resolution $\Delta t$ are determined by the window size and the sampling rate. For style audio, the frequency resolution is approximately 43 Hz, and the time resolution is around 23 ms, whereas for content audio, the frequency resolution improves to approximately 21.5 Hz with a time resolution of about 46 ms.

**Feature Extraction Using the Mel Scale and MFCC for Style and Content Audio**

The feature extraction stage in the music style MA operates distinctly for style and content audio using Mel scale filtering and Mel-Frequency Cepstral Coefficients (MFCC), respectively. Each technique is tuned to highlight specific audio characteristics essential for style transfer and content preservation.

i) Mel Scale Filtering for Style Audio: After obtaining the spectrogram via STFT, the next step involves converting this spectrogram into the Mel scale. The frequencies obtained from the STFT are mapped to the Mel scale using the EQU (3):

$$m = 2595 \log_{10} \left( 1 + \frac{f}{700} \right)$$  \hspace{1cm} (3)

where $m$ is the Mel frequency and $f$ is the linear frequency. A set of triangular filters, each corresponding to a band on the Mel scale, is then applied to the power spectrum of the spectrogram. The filter bank consists of $M$ filters and the output $S_m(k)$ for the $k$-th filter is calculated as EQU (4).

$$S_m(k) = \sum_{n=1}^{N} H_k(f_n) \cdot |X(f_n)|^2$$  \hspace{1cm} (4)
where $H_k(f)$ is the triangular filter function for the $k$-th Mel band, and $|X(f_n)|^2$ is the power at frequency $f_n$ in the spectrogram.

ii) MFCC for Content Audio: Conversely, the content audio utilizes MFCC to capture the broad spectral characteristics fundamental to the audio content. The MFCC extraction involves several steps to transform the Mel-scaled power spectra into cepstral coefficients:

- **Logarithmic Scaling:** The outputs from the Mel filter banks are transformed by taking the logarithm of each filter bank's energy $\log S_m(k)$, this step mimics the non-linear human ear perception of sound, making the features more representative of the perceptual characteristics of the audio.
- **Discrete Cosine Transform (DCT):** A discrete cosine transform is then applied to the log Mel spectra to derive the cepstral coefficients, which provide a compact representation of the audio spectrum, EQU (5).

$$
c_i = \sum_{k=1}^{M} \log S_m(k) \cdot \cos\left[\frac{\pi(k-0.5)}{M}\right]
$$

for $i = 1, 2, ..., L$, where $L$ is the number of cepstral coefficients to be retained, typically much smaller than $M$.

The resulting feature vectors—Mel-filtered spectrogram for style and MFCCs for content—are then processed through their designated convolutional networks, STLCNet and CONCNet, respectively. The feature sets are combined linearly, and their time-frequency representations are shown in Fig. 1.

**Network Model**

STLCNet and CONCNet are structured similarly by having four convolutional layers followed by a fully connected layer. The following section presents in detail about each of these layers:

i) Convolutional Layers:

- **First Layer:** The first convolutional layer applies 32 kernels of size $3 \times 3$ with a stride of $2 \times 2$. Batch normalization is performed to normalize the inputs for each mini-batch, and the Rectified Linear Unit (ReLU) is used as the activation function to introduce non-linearity, EQU (6).
\[ Y_1 = \text{ReLU}(BN(W_1 \ast X + b_1)) \]  

- Second Layer: This layer also uses 32 kernels of size 3 × 3 and a stride of 2 × 2. It includes batch normalization and ReLU activation. Additionally, max-pooling is employed to reduce the spatial dimensions of the feature maps, EQU (7).

\[ Y_2 = \text{ReLU}\left(BN\left(\text{MaxPool}(W_2 \ast Y_1 + b_2)\right)\right) \]

- Third Layer: This layer increases the number of kernels to 64, maintaining the kernel size of 3 × 3 and stride of 2 × 2. Batch normalization and ReLU activation are used further to process the features, EQU (8).

\[ Y_3 = \text{ReLU}\left(BN(W_3 \ast Y_2 + b_3)\right) \]

- Fourth Layer: Similar to the third layer, it uses 64 kernels of size 3 × 3 and stride of 2 × 2, with batch normalization and ReLU activation, EQU (9).

\[ Y_4 = \text{ReLU}\left(BN(W_4 \ast Y_3 + b_4)\right) \]

\textit{ii) Fully Connected Layer:} The processed data from the fourth convolutional layer is fed into a fully connected layer with 1024 hidden units, using the Sigmoid activation function to ensure the network output is bounded and normalized, EQU (10).

\[ Y_5 = \text{Sigmoid}(W_5 \ast Y_4 + b_5) \]

\textit{iii) Output Layer:} This layer transforms the processed features from the previous layers into a matrix that represents the audio’s time-frequency characteristics. The mathematical expression for the output layer is given by EQU (11).

\[ Y_{\text{out}} = W_{\text{out}} \ast Y_5 + b_{\text{out}} \]

In this expression, \( W_{\text{out}} \) denotes the weights of the output layer, \( Y_5 \) represents the activated features from the last fully connected layer and \( b_{\text{out}} \) is the bias term. The resulting output \( Y_{\text{out}} \) is a matrix where each element corresponds to a specific time-frequency component of the audio signal.

Mathematical Theory of Evidence (MTE) Based Fusion

The Framework of Belief Functions (FBF), also recognized as the Mathematical Theory of Evidence, is operational within a discernment space, denoted as \( \Omega \), which comprises a set of hypotheses \( \{H_1, H_2, ..., H_n\} \), where each hypothesis \( H_i \) is an element of the power \( P(\Omega) \). In this setting, a belief mass function, denoted by \( m \), maps the power set \( P(\Omega) \) to the interval \([0,1]\). This function assigns a belief mass to each subset \( H_i \subseteq \Omega \), where:

- The total belief mass distributed among all subsets \( H \) of \( \Omega \) equals 1, expressed as EQU (12).

\[ \sum_{H \subseteq \Omega} m(H) = 1. \]

- The belief mass assigned to the empty set is zero, \( m(\emptyset) = 0 \), to ensure normalization.

In our model, the outputs from the convolutional networks STLCNet and CONCNet, which produce spectrogram feature matrices, are integrated into the belief mass functions \( m_1 \) and \( m_2 \). These functions utilize the detailed features from both networks to perform evidential reasoning about the style and content of audio data. The integration of these belief masses is conducted using Dempster’s Rule of Combination. For any non-empty subset \( A \subseteq \Omega \), the combined belief mass function \( m_{1 \oplus 2}(A) \) is calculated as EQU (13) and EQU (14).

\[ m_{1 \oplus 2}(A) = \frac{1}{1-K} \sum_{BnC=A} m_1(B)m_2(C) \]  

where \( K \) is the normalization factor.
Ethnic Music Style Migration Algorithm Based on Artificial Intelligence

\[ K = \sum_{B \cap C = \emptyset} m_1(B)m_2(C) \]  

(14)

This normalization factor, \( K \), accounts for the sum of products of belief masses from pairs of subsets \( B \) and \( C \) that do not intersect, ensuring the validity of the combined belief mass. The result of this fusion, expressed as the fused spectrogram \( Y_{\text{fused}} \), that encapsulates the integrated features from both style and content networks, EQU (15).

\[ Y_{\text{fused}} = m_1 \oplus_2 \Omega \]  

(15)

Audio Reconstruction

To reconstruct audio from the fused spectrogram, the model uses the Phase Gradient Heap Integration (PGHI) algorithm through the following steps:

Gradient Calculation: The first step in PGHI involves calculating the time-frequency gradient of the phase. The gradients are computed assuming that the valid phase values should produce a consistent and slowly varying time-frequency representation. Mathematically, this is expressed as EQU (16).

\[ \nabla \phi = \left( \frac{\partial \phi}{\partial t}, \frac{\partial \phi}{\partial \omega} \right) \]  

(16)

where \( \phi \) represents the phase, \( t \) is the time index, and \( \omega \) is the frequency index.

Phase Reconstruction: Starting from an initial arbitrary phase (often set to zero), the PGHI uses the calculated gradients to refine the phase estimate iteratively. The iterative process aims to minimize the difference between the magnitude of the original (input) and reconstructed spectrogram. The updated formula for the phase in each iteration is given by EQU (17).

\[ \phi^{(\text{new})} = \phi^{(\text{old})} + \nabla \phi \cdot \Delta t \]  

(17)

where \( \Delta t \) is a small time step, and \( \phi^{(\text{old})} \) is the phase from the previous iteration.

Heap Integration: To ensure a globally consistent phase estimate, PGHI employs a heap-based integration scheme. This scheme prioritizes integrating phase values from regions with higher confidence (typically where the magnitude of the spectrogram is more significant) to regions with lower confidence. This method helps in maintaining the integrity of phase relationships across the spectrogram.

Audio Reconstruction: Once a satisfactory phase spectrum is obtained, it is combined with the magnitude spectrogram. The final reconstructed time-domain signal is obtained by applying the inverse STFT (ISTFT) to the complex spectrogram (constructed using the reconstructed phase and the original magnitude). The ISTFT is defined as EQU (18).

\[ x(n) = \sum_m X(m, \omega) e^{j\phi(m, \omega)} W^{-1}(n - mR) \]  

(18)

where \( X(m, \omega) \) is the magnitude spectrogram, \( \phi(m, \omega) \) is the reconstructed phase, \( W^{-1} \) is the inverse window function, and \( R \) is the hop size.

Algorithm:

\begin{align*}
\text{Input:} & \\
\text{Style audio signal, } s[n] & \\
\text{Content audio signal, } c[n] & \\
\text{Output:} & \\
\text{Fused audio signal incorporating the style of } s[n] \text{ with the content of } c[n] & \\
\text{Parameters:} & \\
\end{align*}

- $N_s, N_c$: Window lengths for style and content audio
- $R_s, R_c$: Hop sizes for style and content audio
- $M$: Number of Mel filters
- $L$: Number of MFCC coefficients

**Procedure:**

1. **Preprocessing:**
   - For both style and content audio:
     - Apply STFT using Hanning windows:
       - **Style:** $X_s(m, \omega) = \sum_n s[n] \cdot w(n - mR_s) \cdot e^{-j\omega n}$
       - **Content:** $X_c(m, \omega) = \sum_n c[n] \cdot w(n - mR_c) \cdot e^{-j\omega n}$

2. **Feature Extraction:**
   - **Style Audio:**
     - Convert spectrogram to Mel scale: $S_m(k) = \sum_n H_k(f_n) \cdot |X_s(f_n)|^2$
   - **Content Audio:**
     - Extract MFCC:
       - Compute Mel-filtered power spectrum and logarithmic scaling.
       - Apply DCT to log Mel spectra to obtain MFCCs: $c_i = \sum_k \log S_m(k) \cdot \cos \left(\frac{\pi(k-0.5)}{M}\right)$

3. **Convolutional Processing:**
   - Process style and content features through respective networks, STLCNet and CONCNet, to obtain feature transformations $Y_{\text{style}}$ and $Y_{\text{content}}$.

4. **Fusion Using Mathematical Theory of Evidence:**
   - Define belief mass functions $m_1$ and $m_2$ based on outputs from STLCNet and CONCNet.
   - Fuse using Dempster's Rule of Combination:
     - $m_1 \oplus_2 (A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B)m_2(C)$
     - Calculate normalization factor $K = \sum_{B \cap C = \emptyset} m_1(B)m_2(C)$

5. **Phase Reconstruction and Audio Synthesis Using PGHI:**
   - Calculate gradients and perform iterative phase reconstruction.
   - Apply ISTFT to synthesize the fused audio from the magnitude and reconstructed phase spectrum.

**EXPERIMENTAL ANALYSIS**

The hardware for the experiment analysis includes an Intel Xeon Processor with at least 16 cores, an NVIDIA Tesla V100 GPU, 64 GB of RAM, and a 2TB SSD. The software includes Linux Ubuntu 20.04, Python 3.8 or newer, TensorFlow 2.x for DL, Librosa for audio feature extraction, and NumPy and SciPy for computations.
The database used for analysis in this work is the Smithsonian Folkways Collection, maintained by the nonprofit record label of the Smithsonian Institution. This extensive collection comprises over 60,000 tracks from various ethnic and cultural groups worldwide, each thoroughly documented with cultural, geographical, and historical details. The tracks are provided in high-quality audio formats and are accompanied by detailed metadata, including information about the artists, instruments, and the cultural significance of the recordings.

The chosen metrics for this evaluation include:

1. Style Transfer Accuracy: Evaluated using KL Divergence to measure how precisely the style characteristics from the style audio are incorporated into the content audio. The EQU (19) for KL divergence is given as:

   \[ D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) \log \left( \frac{P(x)}{Q(x)} \right) \]  

   where \( P \) is the distribution of features in the style audio, and \( Q \) is the distribution in the style-transferred audio. Lower values indicate better style transfer.

2. Content Preservation: Assessed using Signal-to-Distortion Ratio (SDR) to determine how well the content's original characteristics are retained, EQU (20),

   \[ SDR = 10 \cdot \log_{10} \left( \frac{\| s \|^2}{\| s - \hat{s} \|^2} \right) \]  

   where \( s \) is the original content audio signal, and \( \hat{s} \) is the style-transferred audio signal. Higher values indicate better preservation of the original content.

3. Signal-to-Noise Ratio (SNR): Used to evaluate the clarity of the audio by measuring the level of desired signal relative to background noise, EQU (21).

   \[ SNR = 10 \cdot \log_{10} \left( \frac{\text{Power of Signal}}{\text{Power of Noise}} \right) \]  

   Higher values indicate a more precise signal with less background noise.

4. Spectral Convergence: Measures the closeness of the spectral features between the transformed and original content audio. It is done by measuring the Euclidean Distance, EQU (22).

   \[ d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \]  

   where \( p \) and \( q \) are two vectors representing the spectral features of the original and style-transferred audios, respectively. \( p_i \) and \( q_i \) are the components of vectors \( p \) and \( q \). \( n \) is the number of dimensions in each vector, corresponding to the number of spectral bands or features being compared.

5. PEAQ (Perceptual Evaluation of Audio Quality) Provides a standardized auditory test to assess the perceived QoA compared to the original.

6. Fréchet Audio Distance (FAD): Quantifies the similarity between distributions of deep features from original and style-transferred audio, EQU (23).

   \[ FAD = \| \mu_x - \mu_g \|^2 + \text{Tr} \left( \Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{1/2} \right) \]  

   where \( \mu_x, \Sigma_x \) are the mean and covariance of the feature vectors of the original audio and \( \mu_g, \Sigma_g \) are those of the generated audio.

7. Accuracy: Measures the correctness of the style representation in the transformed audio.

8. Mean Opinion Score (MOS): Gathers subjective evaluations from listeners regarding the QoA.

9. Inception Score (IS): Adapted from image processing, this score assesses the clarity and diversity of the generated audio outputs, EQU (24).
where, \( \sim p_g \), \( x \): Represents a generated audio sample, \( p_g \): The probability distribution of the generated audio samples, \( p(y \mid x) \), the conditional probability distribution of labels \( y \) given the generated audio sample \( x \). \( D_{KL}(p(y \mid x) \parallel p(y)) \), The Kullback-Leibler divergence between the conditional probability distribution \( p(y \mid x) \) and the marginal probability distribution \( p(y) \).

To evaluate the algorithm, a set of style-content pairs was selected:

- Pair 1: West African Drumming + Japanese Shamisen
- Pair 2: West African Drumming + Appalachian Folk
- Pair 3: Indian Classical Sitar + Scottish Bagpipe
- Pair 4: Indian Classical Sitar + Andean Pan Flute
- Pair 5: Flamenco Guitar + Native American Chanting
- Pair 6: Flamenco Guitar + Indonesian Gamelan

The results for style transfer accuracy, content preservation, and signal clarity are presented in Fig. 2. Among the pairs tested, Pair 3 (Indian Classical Sitar + Scottish Bagpipe) and Pair 5 (Flamenco Guitar + Native American Chanting) had better outcomes in all metrics. Specifically, Pair 5 exhibited the best results by achieving the lowest KL Divergence (0.25) alongside the highest scores in both Signal-to-Distortion Ratio (20 dB) and Signal-to-Noise Ratio (27 dB). Similarly, the Pair 3 performs better with its high SNR of 25 dB. Conversely, Pair 2 (West African Drumming + Appalachian Folk) shows the weakest performance with the highest KL Divergence (0.50), the lowest SDR (14 dB) and an acceptable SNR of 20 dB. Pairs involving the Indian Classical Sitar (Pairs 3 and 4) generally perform well, which shows that the algorithm effectively handled the spectral and acoustic properties of the sitar when transferring its style to other content types.
The above Fig. 3 displays the performance score for each pair measured for Spectral Convergence, PEAQ Score, and Fréchet Audio. The Spectral Convergence measures how closely the spectral features of the transformed audio align with that of the original content. Lower values indicate a closer match. Pair 5 (Flamenco Guitar + Native American Chanting) stands out with the best spectral convergence score of 0.02, indicating almost perfect spectral alignment. Pair 3 (Indian Classical Sitar + Scottish Bagpipe) also performs exceptionally well, with a score of 0.03.

In contrast, Pair 2 (West African Drumming + Appalachian Folk) has the highest value at 0.07, indicating less effective spectral matching than the other pairs. PEAQ Scores reflect the perceived QoA from a listener's perspective, with higher scores indicating better perceived quality. Pair 5 again scores highest with a PEAQ score of 4.7, aligning with its excellent spectral convergence results, suggesting that not only is the spectral feature match excellent, but it is also perceptually pleasing to listeners. Pair 3 follows closely with a score of 4.5, while Pair 2 has the lowest score at 3.8, suggesting that listeners found the QoA and fidelity less satisfying.

Fréchet Audio Distance (FAD) measures the similarity between the distributions of deep features from the original and transformed audios, with lower scores indicating more significant similarity. Pair 5 has the lowest FAD at 0.55, complementing its top results in the other two metrics and affirming its superior performance in maintaining the integrity and quality of the audio. Pair 3 also shows strong performance with a score of 0.65. Conversely, Pair 2 shows the highest FAD at 0.90, indicating significant differences between the original and transformed audio features, suggesting less effective style transfer and content preservation.
The above Fig. 4 presents the result of the algorithm evaluation using metrics such as Accuracy Score, Mean Opinion Score (MOS), and Inception. Accuracy Score assesses how correctly the algorithm applies the intended style to the content audio. Pair 5 (Flamenco Guitar + Native American Chanting) leads with an outstanding accuracy of 97%. Pair 3 (Indian Classical Sitar + Scottish Bagpipe) also shows excellent accuracy at 95%. In contrast, Pair 2 (West African Drumming + Appalachian Folk) records the lowest accuracy at 88%.

Mean Opinion Score (MOS) measures subjective evaluations from listeners regarding the quality of the audio. Again, Pair 5 excels with the highest MOS of 4.7, corroborating its top performance in technical accuracy and suggesting that listeners find the QoA exceptionally good. Pair 3 follows closely with a MOS of 4.5. The lowest MOS is seen in Pair 2 at 3.8, reflecting listeners’ lower satisfaction with the QoA, which may correlate with its lower accuracy score. Inception Score (IS), adapted for audio, evaluates the clarity and diversity of the generated outputs. Higher scores suggest that the outputs are both diverse and distinguishable. Pair 5 again scores highest at 3.8, indicating high quality and diverse and distinct audio outputs. Pair 3 scores 3.5, demonstrating good diversity and clarity—Pair 2, consistent with other metrics, scores lowest at 2.9.

**CONCLUSION and FUTURE WORK**

Recently, the importance of cultural heritage preservation has been considered and acknowledged globally, and such heritage domains as Ethnic Music (EM) need to be protected and enriched. This research introduced the EM style Migration Algorithm (MA) based on Artificial Intelligence (AI) to maintain EM styles by enabling creative musical interactions across different cultures through the support of the AI model. Using the proposed AI model, the study ensured that the integrity of traditional music is maintained well enough even when integrated with other styles. The algorithm was measured using different evaluation metrics, and it demonstrated effectiveness in style transfer, as indicated by low KL Divergence scores. It also maintained Quality of Audio (QoA) and clarity by showing high SDR and SNR values. Feedback through Mean Opinion Scores also helps confirm the proposed algorithm's capability to produce appealing transformed music. Through these results, the study showcased that the proposed algorithm has acceptable potential for use in music production and cultural exploration.

This work will include automatic music transformation technologies to enhance efficiency further.
REFERENCES


Li, Y., Liu, M., He, F., & Li, L. (2020). Personalized Music Recommendation Based on Style Type.


