IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 2, April, 2024, pp. 1063-1072
Journal homepage: http://iieta.org/journals/ts

Synthesis and Restoration of Traditional Ethnic Musical Instrument Timbres Based on Time-Frequency Analysis

Mengmeng Chen[1], Yu Xiang[1], Chuixiang Xiong[2*]

[1] School of Music and Dance, QuJing Normal University, Qujing 655011, China
[2] School of Continuing Education, Hankou University, Wuhan 430212, China

Corresponding Author Email: Xcx37175809@163.com

https://doi.org/10.18280/ts.410247

Received: 3 December 2023
Revised: 17 February 2024
Accepted: 10 March 2024
Available online: 30 April 2024

Keywords:
traditional ethnic musical instruments, time-frequency analysis, timbre separation, timbre synthesis, timbre restoration, transformer model, signal processing

ABSTRACT

With the advent of the digital age, the preservation and restoration of the timbres of traditional ethnic musical instruments have emerged as significant areas of study in musicology and signal processing. Music serves not only as a bridge between history and culture but also plays an irreplaceable role in expressing ethnic characteristics and emotions. The timbres of traditional ethnic musical instruments, owing to their unique musical expressiveness and cultural value, have attracted widespread attention from both the academic and industrial sectors. However, many valuable timbre recordings are facing threats of damage and disappearance due to limitations in old recording technologies and preservation conditions. Moreover, existing timbre processing technologies still require improvements in separation accuracy, synthesis authenticity, and restoration naturalness. This study aims to achieve efficient separation, authentic synthesis, and natural restoration of the sounds of traditional ethnic musical instruments through advanced signal processing methods. Initially, this paper discusses a sound separation technique for traditional ethnic musical instruments based on time-frequency analysis, addressing the issue of insufficient resolution in complex audio signals. Subsequently, it proposes a timbre synthesis method based on the *Transformer* deep learning model, which can understand and reproduce the delicate timbral characteristics of musical instruments. Finally, addressing the continuity issue in timbre restoration, this paper introduces an innovative restoration technique to enhance the quality of damaged audio restoration and auditory consistency. Through the application of these methods, this study not only contributes to the protection and restoration of traditional timbres but also advances related audio processing technologies.

1. INTRODUCTION

Music, as a cultural carrier that transcends time and space, has always been an important component of human emotional expression and cultural inheritance [1-3]. In particular, the timbres of traditional ethnic musical instruments, which contain rich historical and cultural information and carry the unique charm of national music [4-6], are of significant interest. However, over time, many recordings of traditional instrument timbres face the risk of degradation and loss, presenting severe challenges to the preservation and restoration of these sounds. To protect these precious musical heritages, research on the separation, synthesis, and restoration of traditional ethnic musical instrument sounds based on time-frequency analysis is particularly important [7].

Currently, research on the timbres of traditional ethnic musical instruments is not only beneficial for the digital preservation and inheritance of music but also provides new application scenarios for the advancement of audio processing technology [8, 9]. Through accurate timbre separation and synthesis, we can reproduce the pure sounds of ethnic instruments in different musical works, and even reuse these unique sound materials in modern music creation [10-12]. Moreover, timbre restoration technology plays a vital role in the recovery of damaged audio, allowing historical audio materials to be revived and supporting research in music culture.

Despite numerous studies in the fields of timbre separation, synthesis, and restoration, existing methods are often limited by the resolution constraints of time-frequency analysis, making it difficult to handle complex audio signals, especially the separation of mixed sounds from traditional ethnic musical instruments [13, 14]. Furthermore, existing synthesis methods often lack sufficient flexibility and expressiveness when dealing with the subtle differences in the timbres of ethnic instruments, while timbre restoration techniques still have shortcomings in continuity processing and naturalness restoration [15-17].

In response to these issues, this paper first proposes a sound separation technique for traditional ethnic musical instruments based on time-frequency analysis, which can effectively extract the target instrument sounds from mixed signals. Secondly, the paper introduces a timbre synthesis method based on the *Transformer* model, capable of capturing and generating highly realistic ethnic instrument timbres. Finally, this paper also explores a novel continuity-based timbre restoration technique, aimed at improving the naturalness and coherence in the restoration of damaged audio by traditional timbre restoration methods. These studies not only provide new technical means for the digital protection and restoration of traditional ethnic musical instrument timbres but also bring potential application value to fields such as music information retrieval and automatic accompaniment generation.

2. TRADITIONAL ETHNIC MUSICAL INSTRUMENT SOUND SEPARATION BASED ON TIME-FREQUENCY ANALYSIS

Non-negative Matrix Factorization (*NMF*) is a technique commonly used in audio signal processing, especially in source separation tasks. Through *NMF*, the time-frequency representation of an audio signal can be decomposed into fundamental spectral patterns and corresponding activation coefficients. Long windows and short windows have different characteristics in time-frequency analysis. Long window spectrograms are more suitable for analyzing and separating sound components that are continuous and change slowly, such as sustained notes or continuously played instruments (e.g., pipe organ, sustained bowing sounds of string instruments). In traditional ethnic musical instruments, if the timbre and pitch of the instrument are relatively stable, long window analysis will be more effective. Analyzing with shorter time windows can achieve higher time resolution but sacrifices frequency resolution. Short windows are better suited for capturing rapidly changing sounds, such as the strike of percussion instruments or the plucking of guitar strings. For traditional ethnic musical instruments, short window analysis can better handle components that contain rapid rhythms or abrupt changes in pitch or timbre.

2.1 Attenuation based on long window spectrogram decomposition

In long window spectrograms, the sound characteristics of fast rhythm parts are significantly different from those of percussion instruments and vocals. For fast rhythms, especially those composed of rapidly alternating notes, they usually appear as rapid changes along the time axis on the spectrogram. Since each note has a short duration, these changes form a series of brief and clear stripes along the time axis, in contrast to the wide-band continuity of percussion instruments and the frequency dispersion of vocals, the stripes of the fast rhythm parts are more slender and dispersed. Transient acoustic events, such as the plucking sound of an instrument or a strike, refer to sounds that appear briefly in musical performance but have significant energy peaks. In long window spectrograms, these transient events are represented by sharp peaks on the time axis and wider bands on the frequency axis. Compared to the sound of percussion instruments, transient events usually have higher energy and shorter duration, making them appear as more distinct instant peaks on the spectrogram. Relative to the continuous energy distribution of vocals, although transient events also show some continuity on the frequency axis, their concentration of energy and the brief nature of their duration make them distinctly different from vocals.

In this study, to filter out the sounds of fast rhythm sections and transient acoustic events in music, an innovative method based on *NMF* is adopted, as illustrated in Figure 1. Initially, the input music undergoes a short-time Fourier transform (*STFT*) with a long window to construct an amplitude spectrogram that does not contain phase information. This step allows for the analysis of the music's time-frequency characteristics without interference from phase changes. Then, by decomposing the amplitude spectrogram using *NMF*, multiple *NMF* components can be obtained, each representing different sound components in the music signal. Since the characteristics of fast rhythms and transient acoustic events in long window spectrograms are discontinuities along the frequency axis, the key strategy of this paper is to identify those *NMF* components that exhibit frequency discontinuities and remove them from the music mix. This process is essentially a component filtering within the time-frequency domain, effectively removing or reducing components considered to be fast rhythms and transient acoustic events, while retaining other continuous sound components. The goal of this method is to purify specific sound events in music without compromising the overall structure and texture of the music, providing purer sound materials for the synthesis and restoration of traditional ethnic musical instrument timbres.

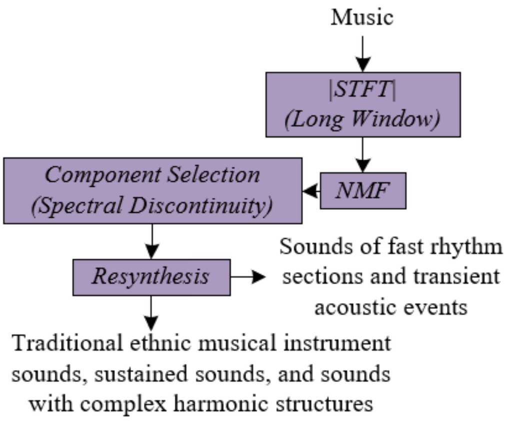

Music
|STFT|
(Long Window)
NMF
Component Selection
(Spectral Discontinuity)
Resynthesis
Sounds of fast rhythm sections and transient acoustic events
Traditional ethnic musical instrument sounds, sustained sounds, and sounds with complex harmonic structures

Figure 1. Method for attenuating fast rhythm sections and transient acoustic events

An important step in synthesizing and restoring the timbres of traditional ethnic musical instruments is to remove noises that may mask or affect the essential timbre of the instruments, such as fast rhythm sections and transient acoustic events. The spectral discontinuities these sound components exhibit in the spectrogram may contrast with the continuous and harmonious characteristics of instrument timbres, interfering with the extraction and analysis of pure instrument timbres. Therefore, it is crucial to devise a method to evaluate the spectral discontinuity of components after *NMF* decomposition, to selectively delete or retain specific *NMF* components. For this purpose, a method for determining the value of spectral discontinuity was designed. Specifically, this method involves quantitatively assessing the spectral discontinuity of each *NMF* component. This may involve calculating the variation of components across the frequency dimension, identifying those with significant energy changes between adjacent frequency points, as these changes signify discontinuity. By computing the difference or gradient of each row of the *NMF* base matrix, the rate of change of the fundamental components across the frequency dimension can be determined, thereby setting a threshold to decide whether a component is considered continuous or discontinuous. Components with higher continuity are retained, as they may relate to the sustained timbre of instruments, while those with higher discontinuity, thought to be produced by fast rhythms or transient acoustic events, should be removed from the synthesized timbre.

The specific k-th (k=1,...,K) *NMF* component is A^k, composed of the k-th column of Y and the k-th row of H. A^k's degree of spectral discontinuity is small, represented by $f_t(A^k)$. Assuming the k-th column of Y is represented by y_k, and the normalization term is represented by $\sum^{J}_{j=1}y^2_k(j)$, the formula for calculation is provided below:

$$f_t\left(A^k\right)=\frac{\sum_{j=2}^{J}\left(y_k(j)-y_k(j-1)\right)^2}{\sum_{j=1}^{J}y_k^2(j)} \tag{1}$$

The higher the value of $f_t(A^k)$, the greater the discontinuity of A^k on the spectrum. Assuming the threshold is represented by ϕ_t, and the *NMF* components belonging to fast rhythm sections and transient acoustic events are represented by A^k, then $f_t(A^k)$ satisfies the inequality shown below:

$$f_t\left(A^k\right)>\varphi_t \tag{2}$$

After identifying all components that satisfy the above inequality representing fast rhythm sections and transient acoustic events, these components are further removed to obtain a new amplitude spectrogram A'. Assuming 0 is a zero matrix of the same size as A, the k-th column of Y is represented by y_k, the k-th row of H is represented by h_k, and the element-wise maximum between X and Y is represented by $MAX(X,Y)$. The expression is as follows:

$$A' = MAX\left(0, A-\sum_{\substack{k=1,\dots,K\\ f_t\left(A^k\right)>\varphi_t}} y_k h_k\right) \tag{3}$$

2.2 Attenuation based on short window spectrogram decomposition

In short window spectrograms, sustained sounds usually exhibit continuity on the time axis because their energy is evenly distributed over time, while they may appear as discrete multiple peaks on the frequency axis, reflecting their harmonic composition. Sound parts with a complex harmonic structure, due to lower frequency resolution in short window spectrograms, become less distinct; their rich harmonics may be merged into wider frequency bands, losing some detail but still maintaining continuous distribution characteristics on the frequency axis. This is similar to the sound of harmonic instruments, as the timbre of harmonic instruments is also composed of the fundamental and a series of harmonics, which show continuity on the frequency axis. However, compared to traditional ethnic musical instrument sounds, the frequency continuity of sustained sounds and sounds with complex harmonic structures may be more apparent because many ethnic instrument sounds contain unique non-linear harmonic components or ornaments, which may exhibit more complex and irregular patterns in short window spectrograms. Compared to percussion instruments, these sounds have obvious temporal continuity in the time-frequency graph, rather than the transient, impulsive energy distribution characteristic of percussion instruments.

Inspired by the short window spectrogram's ability to discriminate sound properties, this paper proposes a method using short window *STFT* to capture the transient energy distribution in music signals, specifically for attenuating sustained sounds and sounds with complex harmonic structures in traditional ethnic musical instrument music. Although this method sacrifices frequency resolution, it better retains details in time, making transient and sustained sounds easier to distinguish. During the *NMF* decomposition process, this paper specifically focuses on those components that show continuity on the time axis and discrete distribution on the frequency axis, usually corresponding to sustained sounds and sounds with a complex harmonic structure. Through a carefully designed component selection strategy, this paper can identify and attenuate these specific components from complex audio mixes, thereby reducing or removing unwanted sounds, such as accompaniment or other interference elements, to highlight and preserve the core timbral characteristics of the instrument. The principle flow is shown in Figure 2.

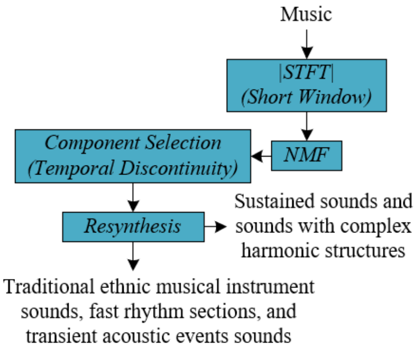

Music
|STFT|
(Short Window)
NMF
Component Selection
(Temporal Discontinuity)
Resynthesis
Sustained sounds and sounds with complex harmonic structures
Traditional ethnic musical instrument sounds, fast rhythm sections, and transient acoustic events sounds

Figure 2. Method for attenuating sustained sounds and sounds with complex harmonic structure

Specifically, an algorithm is created to analyze the temporal characteristics of components in the short window spectrogram. This algorithm first decomposes the audio signal using *NMF*, then evaluates the temporal discontinuity of each *NMF* component by calculating its temporal change. This can be achieved through methods such as counting the number of energy peaks in a short time, the frequency of energy surges, or using a custom time-varying curve fitting error. Once the component's temporal discontinuity is quantified, a threshold can be set, with components exceeding this threshold identified as percussion sounds and removed. Determining this threshold may rely on analyzing a large number of samples to find the optimal separation effect or may introduce machine learning techniques to automatically adjust the threshold, thereby adapting to different music and recording conditions.

Assuming the k-th row of H is represented by h_k, and the normalization term is represented by $\sum^{S}_{s=1}h^2_k(s)$, the temporal discontinuity degree of the k-th NMF component A^k is represented by $f_s(A^k)$, with the calculation formula provided below:

$$f_s\left(A^k\right)=\frac{\sum_{s=2}^{S}\left(h_k(s)-h_k(s-1)\right)^2}{\sum_{s=1}^{S}h_k^{\,2}(s)} \tag{4}$$

Generally, the greater $f_s(A^k)$, the more significant the discontinuity of A^k on the time axis. Assuming the threshold is represented by ϕ_s, if $f_s(A^k)$ satisfies the following, it can be considered as an *NMF* component of sustained sounds and sounds with a complex harmonic structure:

$$f_s\left(A^k\right) > \varphi_s \tag{5}$$

3. SYNTHESIS OF TRADITIONAL ETHNIC MUSICAL INSTRUMENT TIMBRES BASED ON TRANSFORMER

To improve the accuracy and efficiency of traditional ethnic musical instrument timbre synthesis, this paper opts to construct a *FastSpeech2*-Generative Adversarial Networks (*GAN*) timbre synthesis model. This model utilizes the *Transformer* architecture as the core framework for the Mel spectrogram generator, leveraging its parallel processing capability to significantly enhance the training and inference speed of timbre synthesis, overcoming the slowness issues associated with previous *RNN*-based structures. The *Duration Predictor* predicts the duration each phoneme should last, combined with a *Length Regulator* to adjust the output of the decoder, ensuring the generated audio matches the target duration, thereby improving the accuracy of phoneme duration prediction and enhancing the naturalness of the synthesized audio. On this basis, *FastSpeech2-GAN* integrates the *PostNet* structure from *Tacotron2* to fine-tune the initially generated Mel spectrogram, further enhancing the quality of the synthesized timbre. Finally, the model employs the training method of *GAN* to optimize the parameters of the *Transformer* generator, to simulate and learn the complex timbre characteristics of traditional ethnic musical instruments, ensuring the authenticity and expressiveness of timbre synthesis, meeting the high fidelity requirements for traditional timbre synthesis and restoration. Figure 3 shows the generator model structure.

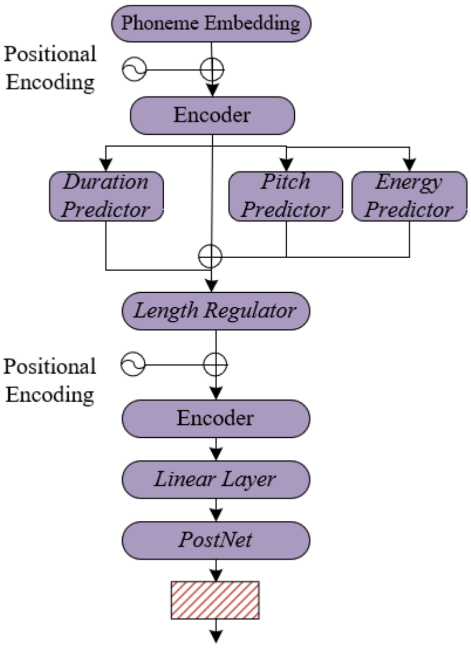


Figure 3. Generator model structure

3.1 Generator

In non-autoregressive speech synthesis systems, accurately predicting the duration of phonemes is one of the crucial factors for achieving naturally sounding speech. Traditional autoregressive speech synthesis systems generate speech incrementally, relying on the output of the previous time step to predict the next, resulting in a slow synthesis process. Additionally, since the actual duration of phonemes in speech varies, relying solely on the encoder's fixed output is insufficient for generating natural speech. To achieve non-autoregressive synthesis of traditional ethnic musical instrument timbres, this paper incorporates *Duration Predictor* and *Length Regulator* modules in *FastSpeech2-GAN*, as shown in Figure 4.

The *Duration Predictor* is a key component responsible for predicting the duration required for each phoneme. *FastSpeech2-GAN* uses a *Duration Predictor* based on one-dimensional convolution and Layer Normalization (*LN*) to directly predict the duration of each phoneme, avoiding the need for step-by-step prediction and significantly increasing the efficiency of speech synthesis. The *Length Regulator* module uses the prediction results of the *Duration Predictor* to adjust the sequence length of the encoder output. The *Length Regulator* extends the encoding of each phoneme through repetition, matching each phoneme's representation to its predicted duration, compensating for the insufficient length of encoder output. This mechanism ensures that the output sequence maintains consistency with the target speech's rhythm and speed in the temporal dimension.

In the text-to-speech (*TTS*) conversion process, the same text fragment can correspond to speech outputs of different loudness, and this variation in vocal energy is crucial for generating naturally sounding speech, as it affects the speech's emotion and emphasis. Therefore, this paper further introduces an *Energy Predictor* module to simulate the dynamic changes in vocal energy, allowing the synthesized speech to better express different speech intensities and emotional nuances. To achieve this, the structure of the *Duration Predictor* is used to construct the *Energy Predictor*, ensuring the model can effectively learn the mapping from text to energy levels. During training, the mean squared error (*MSE*) between the output of the *Energy Predictor* and real energy data is used as a loss function to guide the optimization of model parameters, thereby improving the accuracy of energy prediction. In tonal languages, such as Chinese, accurate prediction of pitch is especially important for generating naturally sounding speech, as pitch variations often carry semantic and grammatical information. For this reason, this paper further introduces a *Pitch Predictor* component responsible for predicting the pitch information of speech. Pitch, the fundamental frequency of speech, is the primary factor determining the pronunciation's tone. The *Pitch Predictor* also adopts a structure similar to the *Duration Predictor* to learn the complex relationship between text and pitch. In this way, the model can generate speech with richer melody and more natural pitch variations. During the training phase, the performance of the *Pitch Predictor* is also optimized by calculating the *MSE* between its output and real pitch data, ensuring the accuracy of pitch prediction. Assuming the predicted and actual values by the *Predictor* are represented by *b*^ and *b*, the *MSE* loss calculation formula is given as follows:

$$MSE = \frac{1}{l}\sum_{u=1}^{l}\left(b_u - \hat{b}_u\right)^2 \tag{6}$$

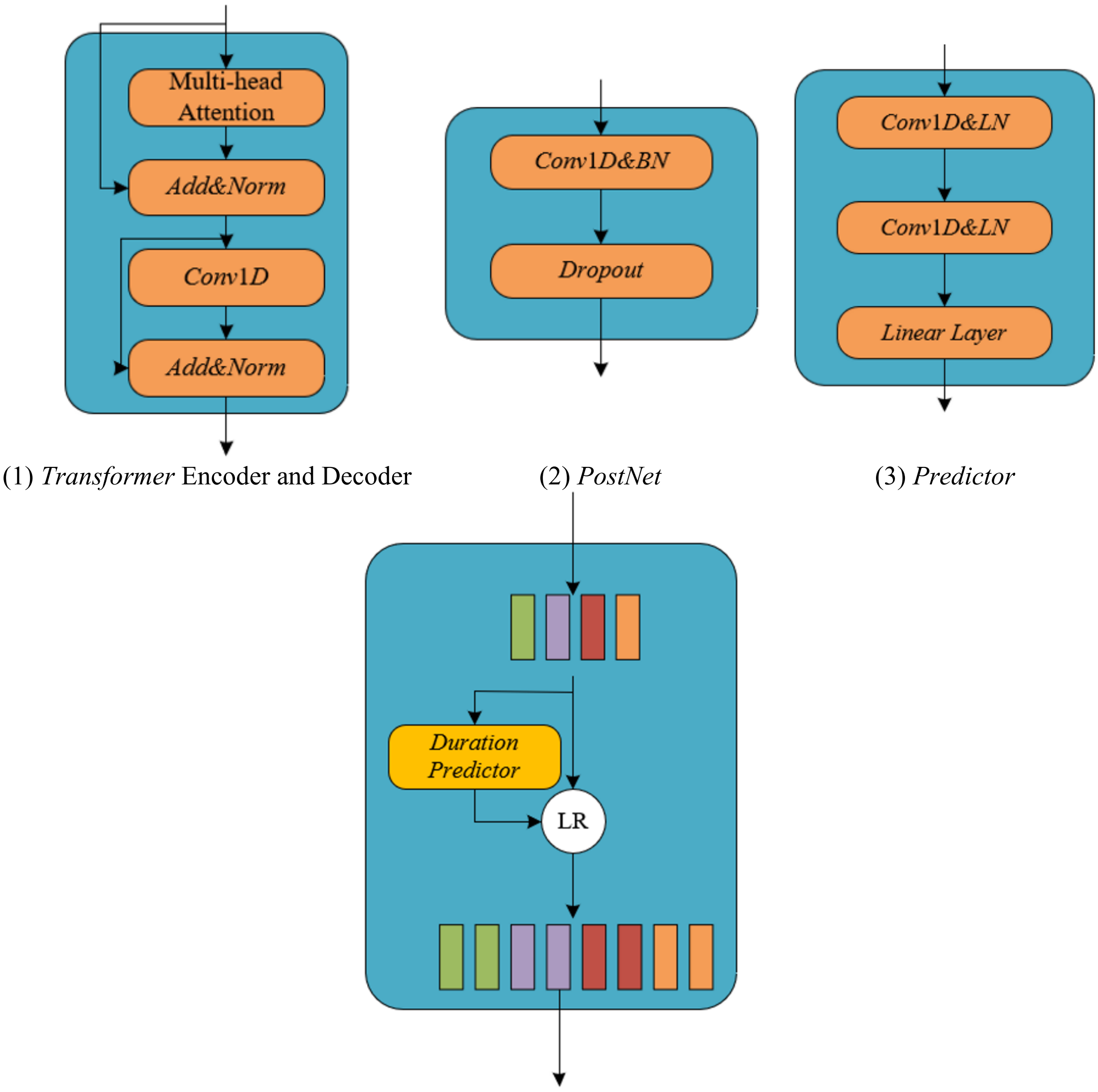
Given that subtle differences in instrument timbres significantly affect the authenticity and recognition of timbres, this paper incorporates the *PostNet* structure from *Tacotron2* to improve and optimize the final synthesized speech quality. *PostNet* is a post-processing network that follows the basic

Mel spectrogram generation part of the model, aiming to refine the initially synthesized Mel spectrogram, making the generated speech more natural and realistic, as shown in Figure 4. *PostNet* typically consists of a series of convolutional layers that can capture details possibly missed by the base spectrum and correct them. By post-processing the Mel spectrogram, *PostNet* helps *FastSpeech2-GAN* more accurately reconstruct the timbral characteristics of instruments, including but not limited to resonance and overtones, thereby enhancing the quality of speech synthesis and providing higher fidelity outputs for the synthesis and restoration of traditional ethnic musical instrument timbres. During training, in addition to the *Predictor*'s *MSE* loss, the model also needs to calculate the Mean Absolute Error (*MAE*) loss between the ME_{BE} before *PostNet* input and the ME_{AF} after *PostNet* output against the real Mel spectrogram features, with the calculation formula given as follows:

$$MAE = \frac{1}{l}\sum_{u=1}^{l} | b_u - \hat{b}_u | \tag{7}$$

Assuming the predicted phoneme duration, vocal energy information, and pitch information are represented by f, r, d, and the actual phoneme duration, vocal energy information, and pitch information are represented by f_{HS}, r_{HS}, d_{HS}. The total generator loss M_H is calculated as follows:

$$\begin{aligned} M_H &= MSE\left(f_{HS}, f\right) + MSE\left(r_{HS}, r\right) + MSE\left(d_{HS}, d\right) \\ &+ MAE\left(ME_{HS}, ME_{BF}\right) + MAE\left(ME_{HS}, ME_{AF}\right) \end{aligned} \tag{8}$$



(1) *Transformer* Encoder and Decoder

(2) *PostNet*

(3) *Predictor*

(4) *Length Regulator*

Figure 4. Advanced module structure introduced by the discriminator

3.2 Discriminator

In the study of traditional ethnic musical instrument timbre synthesis and restoration, authenticity is particularly important, as the timbre of an instrument is required to accurately reproduce fundamental attributes like pitch and loudness, as well as capture unique sound quality details such as resonance characteristics and overtone structures. Under the *GAN* framework, the generator is responsible for creating high-fidelity timbres, while the discriminator evaluates the distinction between synthetic and real timbres. Because the adversarial mechanism of *GAN* can greatly improve the realism and naturalness of speech synthesis, this paper chooses to use the *GAN* training method for model parameter optimization. Through this adversarial training, the generator learns how to produce increasingly high-quality timbres that the discriminator finds difficult to distinguish, thus constantly approaching the real ethnic musical instrument timbres. Such a training mechanism ensures the richness and complexity of the synthesized timbre details, making the restored and synthesized ethnic musical instrument audio not only technically high standard but also audibly more realistic and appealing, meeting the study's goals and quality requirements. Figure 5 shows the discriminator model structure.

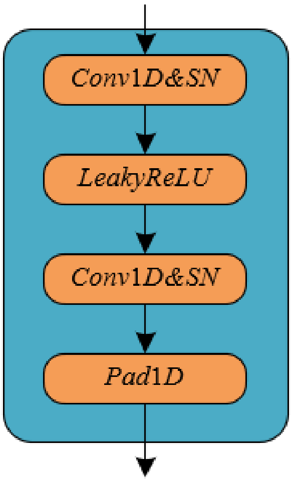


Figure 5. Discriminator model structure

To enhance the stability of the training process in *GAN*, the discriminator of the *FastSpeech2-GAN* model employs spectral normalization instead of the conventional layer normalization. Spectral normalization controls the spectral norm of the weight matrix, effectively constraining the discriminator function's *Lipschitz* constant, which helps to mitigate the mode collapse problem in *GAN* training, ensuring the discriminator is not overly sensitive to the generator's outputs. For complex ethnic musical instrument timbre characteristics, this stability is particularly important as it ensures the model can smoothly capture and differentiate the subtle differences between real and synthetic timbres during the learning process without gradient explosion or disappearance due to training instability. Assuming *L2* regularization is represented by $\|.\|_2$, a and a' are any two values within the domain, the minimum L that satisfies the condition is called the *Lipschitz* constant, and the *Lipschitz* constraint requires the function $d(\cdot)$ to satisfy the following within its domain:

$$\frac{\left\|d(a)-d(a')\right\|_2}{\left\|a-a'\right\|_2} \le L \tag{9}$$

Assuming the *FastSpeech2-GAN* generator is represented by H, the *FastSpeech2-GAN* discriminator by F, the real data distribution by O_{DA}, and the data distribution generated by the generator by O_H, a in the above formula respectively satisfies O_{DA} and O_H. The final optimization goal of the *GAN* during the training process is given by the following formula:

$$\begin{aligned} N(H,F) &= R_{a\sim O_{DA}}\left[\log F(a)\right]+R_{a\sim o_H}\left[\log\left(1\text{-}F(a)\right)\right] \\ &= \int_a O_{DA}(a)\log F(a)\, fa+\int_a O_H(a)\log\left(1\text{-}F(a)\right) fa \\ &= \int_a \left[O_{DA}(a)\log F(a)+O_H(a)\log\left(1\text{-}F(a)\right)\right] fa \end{aligned} \tag{10}$$

4. CONTINUITY-BASED RESTORATION OF TRADITIONAL ETHNIC MUSICAL INSTRUMENT TIMBRES

The timbres of ethnic instruments contain a rich composition of overtones and dynamic changes, necessitating that interpolation methods not only precisely pass through known sample points, i.e., captured timbre characteristics points, but also smoothly transition between sample points to preserve the natural continuity of the timbre. Therefore, this paper chooses cubic spline interpolation for the restoration of traditional ethnic musical instrument timbres. Compared to higher-order polynomial interpolation, cubic spline interpolation avoids the Runge phenomenon, thus ensuring global smoothness while maintaining local control. This method is well-suited to reconstructing subtle fluctuations and details in the timbre signal, which is urgently needed for the restoration of traditional ethnic musical instrument timbres.

Before restoring the timbres of traditional ethnic musical instruments, the primary task is to collect timbre data. This process involves recording the sound produced by the instrument and converting it into a digital signal form, typically waveform data, which contains information about pitch, volume, timbre, and duration. The selection of data points needs to be fine enough to ensure that the characteristics of the instrument timbre are captured. These data points are the nodes (a_u, b_u) mentioned above, representing the time a_u and the corresponding audio intensity or spectral characteristics b_u. The choice of nodes should represent key change points in the timbre signal, such as peaks, troughs, or transition points, so that the cubic spline interpolation can more accurately simulate the dynamic changes of the original timbre.

Once the nodes are determined, the next step is to construct a cubic polynomial function between each pair of adjacent nodes, the collection of these functions forms the piecewise-defined spline curve $T(a)$. For each interval (a_u, a_{u+1}), a cubic polynomial $T_u(a)$ is determined, which not only equals b_u at node a_u but also requires its first and second derivatives at the endpoints to match those of adjacent polynomials, ensuring the continuity of the curve and its derivatives over the entire interval. Specifically, this involves solving a set of linear equations to find the polynomial coefficients, composed of interpolation conditions and smoothness conditions. This process may involve solving a system of linear equations or using specific algorithms (such as interpolation or least squares method) to determine these coefficients.

Finally, the constructed cubic spline function is used to restore missing or damaged timbre data. For each missing area in the timbre signal, the cubic spline polynomial for the corresponding interval can be used to estimate the missing data points and fill in these gaps. Since cubic spline interpolation ensures the smoothness and continuity of the entire timbre curve, the restored timbre signal will be very close to the real instrument output, both in subtle changes within the dynamic range and in the overall texture of the timbre. Ultimately, the timbre data restored and reconstructed through this method should retain the original sound quality characteristics and expressiveness of the instrument as much as possible without introducing artificial distortions.

Given v+1 data points, with v intervals, the cubic spline equation satisfies the following conditions:

(1) In each segmented interval $[a_u,a_{u+1}](u=0,1,.,v-1)$, $T(a)=T_u(a)$ is a cubic polynomial;

(2) Satisfies $T_u(a)=b_u,(u=0,1,\ldots v-1)$;

(3) Ensures that $T(a)$'s first derivative $T'(a)$ and second derivative $T''(a)$ are continuous within the $[x,y]$ interval, meaning $T(a)$'s curve is smooth.

5. EXPERIMENTAL RESULTS AND ANALYSIS

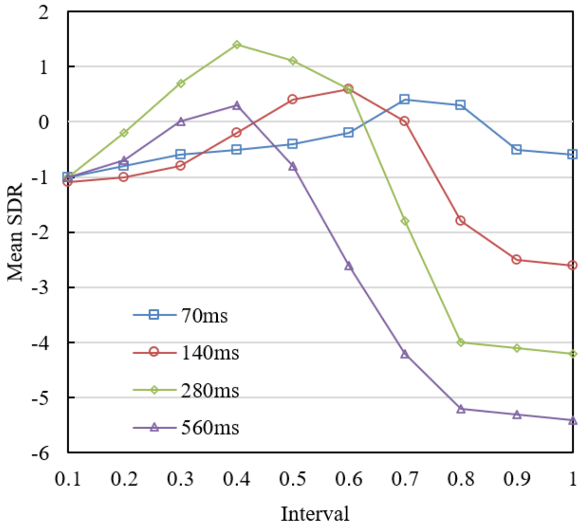


Figure 6. Changes in long window spectrogram decomposition performance under different window lengths

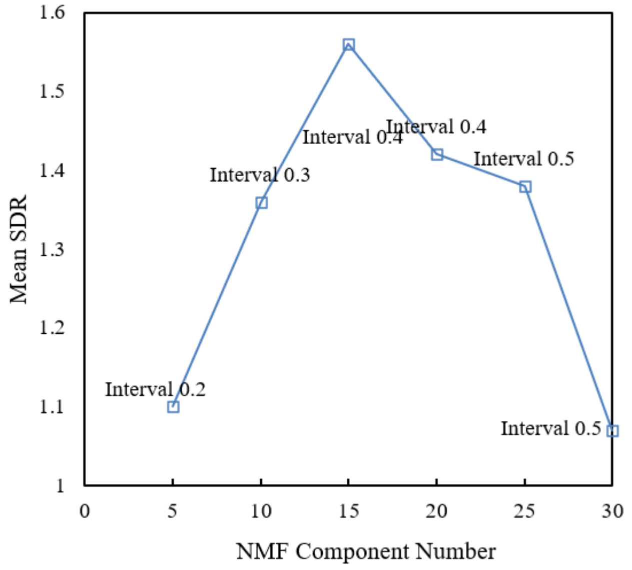


Figure 7. Changes in short window spectrogram decomposition performance under different NMF component numbers

Figure 6 shows the performance changes in spectrogram decomposition of sound signals using different window lengths (70ms, 140ms, 280ms, and 560ms) under various threshold settings. In time-frequency analysis, the window length is a crucial parameter determining time resolution and frequency resolution. At shorter window lengths (70ms and 140ms), the decomposition performance initially worsens (scores becoming more negative) as the threshold increases, then improves at certain thresholds (score increment becoming closer to 0 or positive). This indicates that for shorter windows, the decomposition performance significantly changes with the threshold, related to the increased time resolution, allowing for more refined processing of transient signal changes. For longer window lengths (280ms and 560ms), we observe better performance at lower thresholds (close to zero or positive values) and a sharp decline in performance at higher thresholds (more negative scores). This trend is due to the advantage of long windows in frequency resolution, which better captures stable frequency components, but filtering out some important local time features at higher thresholds leads to performance degradation. Across all window lengths, when the threshold is around 0.5, the performance shows a more balanced result. This suggests that around this threshold, time-frequency analysis achieves a better balance between time and frequency resolution, capturing frequency changes while maintaining sensitivity to time details.

According to the data provided in Figure 7, it can be observed that as the number of *NMF* components increases, the *Mean* Signal-to-Distortion Ratio (*SDR*) initially increases, indicating an improvement in decomposition quality. Specifically, when the number of *NMF* components increases to 10, the *Mean SDR* rises from 1.1 to 1.56. This indicates that with more components used in the decomposition process, the model can more accurately capture and reconstruct the signal's structure. However, after the number of *NMF* components increases to 15, the *Mean SDR* peaks and then begins to decrease, dropping from 1.56 to 1.38, and further decreasing to 1.07 as the components increase to 25 and 30. This phenomenon suggests that although increasing the number of components generally helps to better represent the data, too many components can lead to overfitting or increased model complexity and do not necessarily result in better decomposition outcomes. At 10 *NMF* components, we achieve the highest *Mean SDR* value, indicating that the model has found a balance between accurately decomposing the signal and maintaining model complexity.

From these results, it can be concluded that the time-frequency analysis-based technique for separating traditional ethnic musical instrument sounds proposed in this paper can effectively improve the *SDR*, thereby better extracting the target instrument's sound from mixed signals under an appropriate number of *NMF* components. This demonstrates that the proposed technique achieves effective sound separation, reaching optimal decomposition performance when reasonably selecting the number of components.

Table 1. Iterative synthesis effects of the timbre synthesis model at different training steps

	MOS	*MCD*	*FDSD*	*eFDSD*
Original Music	4.32±0.06	-	-	-
1*k* steps	3.89±0.07	2.8784	0.0158	0.0012
10*k* steps	3.78±0.08	2.9125	0.0158	0.0013
100*k* steps	3.89±0.08	2.8995	0.0157	0.0014

Table 1 shows the iterative synthesis effects of the timbre synthesis model at different training steps, using Mean Opinion Score (*MOS*), Mel Cepstral Distortion (*MCD*), Frame-level Discrete Spectral Distortion (*FDSD*), and enhanced Frame-level Discrete Spectral Distortion (*eFDSD*) as evaluation metrics. Starting with the evaluation of original music, the initial *MOS* value of 4.32 provides a baseline for comparing the quality of synthesized timbres. With the model trained for 1*k* steps, *MOS* decreases to 3.89, and further decreases to 3.78 at 10*k* steps, indicating a decline in the subjective evaluation of synthesized timbres. However, after 100*k* steps of training, *MOS* rises back to 3.89, indicating a recovery in synthesis quality. Although there is slight fluctuation in *MCD*, *FDSD*, and *eFDSD* metrics with increasing training steps, the overall change is minimal, indicating the model maintains relatively stable performance in spectral distortion. These results indicate that the *Transformer*-based timbre synthesis method can approach the quality of original music in terms of synthesized sound quality after sufficient training iterations, as evidenced by the recovery in *MOS*. Although there is a decline in the synthesis performance of the model in the mid-training phase, due to some temporary fitting fluctuations experienced during the learning process, the model ultimately learns the timbre characteristics better and stabilizes its synthesis effect at a level close to the original music. The relative stability of the *MCD*, *FDSD*, and *eFDSD* metrics further indicates the model's consistency and reliability in capturing and reconstructing the spectral details of audio.

Table 2. Synthesis effects of different timbre synthesis models

	MOS	***MCD***	***FDSD***	***eFDSD***
Original Music	4.32±0.06	-	-	-
Attention RNN	3.78±0.08	2.8586	0.0512	0.0187
*FastSpecch*2	3.85±0.09	2.9125	0.0158	0.0017
The Proposed Model	3.89±0.08	2.8995	0.0159	0.0014

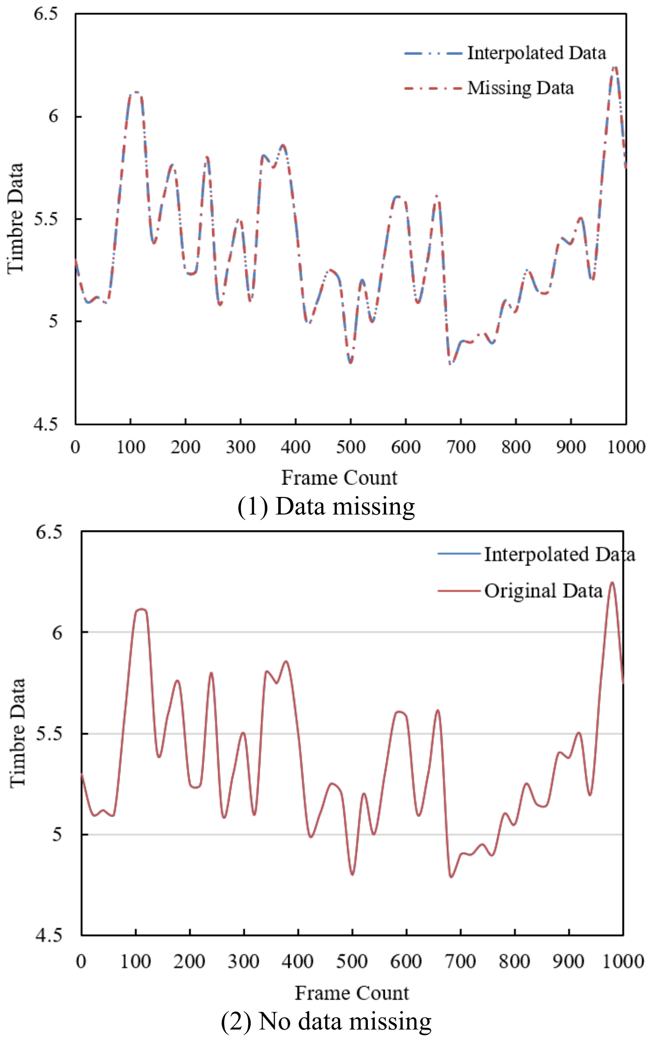
Table 2 shows the synthesis effects of different timbre synthesis models compared to original music. In terms of *MOS*, original music scores the highest at 4.32, providing a reference standard. The *Attention RNN* model has an *MOS* of 3.78, while the *FastSpeech2* model scores slightly higher at 3.85. The *Transformer*-based model proposed in this paper achieves an *MOS* score of 3.89, the closest to the score of the original music among the three models. In terms of *MCD*, the paper's model also shows performance comparable to *FastSpeech2* but slightly better than the *Attention RNN* model. For *FDSD* and *eFDSD* metrics, the paper's model performs similarly to *FastSpeech2* and significantly better than the *Attention RNN* model. These quantitative metrics indicate that the model proposed in this paper has advantages in fidelity and restoration of spectral details of sound quality.

Table 3. Comparison of speech synthesis speeds of different timbre synthesis models

	Training Duration	***RTF***	**Inference Speed Increment**
Attention RNN	200*h*	83.5×10^{-3}	-
*FastSpecch*2	16*h*	11.2×10^{-3}	8*x*
The Proposed Model	18*h*	11.2×10^{-3}	8*x*

Table 3 provides a comparison of different timbre synthesis models in terms of training duration, Real-Time Factor (*RTF*), and inference speed increment. Regarding training duration, the *Attention RNN* model requires 200 hours, while the *FastSpeech2* and the *Transformer*-based model proposed in this paper significantly reduce training time to 16 and 18 hours, respectively. In terms of *RTF*, namely the ratio of the time required to synthesize 1 second of audio to 1 second, both the proposed model and *FastSpeech2* have an *RTF* of 11.2×10^{-3}, much lower than the *Attention RNN* model's *RTF* of 83.5×10^{-3}. In terms of inference speed increment, both the proposed model and *FastSpeech2* achieve an 8-fold speed increase compared to the *Attention RNN* model. These data highlight the advantages of the *Transformer*-based model in training efficiency and high-speed performance during inference.

In summary, the *Transformer*-based timbre synthesis model proposed in this paper not only closely matches the quality of original music in terms of sound quality but also has significant advantages in training efficiency and inference speed. Compared to the traditional *Attention RNN* model, the proposed model greatly reduces training time and maintains an extremely low *RTF* during audio synthesis, ensuring rapid response and efficient processing. Moreover, compared to the *FastSpeech2* model, the proposed model provides higher quality timbre synthesis effects while maintaining the same inference speed.



(1) Data missing

(2) No data missing

Figure 8. The restoration effect of cubic spline functions on traditional ethnic musical instrument timbre data

Figure 8 displays a comparison between the data interpolated through cubic spline functions and the original timbre data. Observing the two sets of data, it can be seen that the interpolated data closely matches the original data across various frames, indicating that the interpolation process can effectively recover damaged or missing data points. Even in sections where the data fluctuates more, the interpolated data successfully simulates the trends and changes of the original data without noticeable deviations or unnatural fluctuations. This demonstrates that the interpolation method can effectively capture and reproduce the subtle variations in the timbre of ethnic instruments while maintaining the naturalness and coherence of the timbre. From the above analysis, it can be concluded that the continuity-based timbre restoration technique discussed in this paper is highly effective in processing traditional ethnic musical instrument timbre data, regardless of whether the data is missing or intact. Particularly in the restoration of audio signals, continuity is one of the key factors in maintaining the naturalness of sound quality. The technique proposed in this paper is clearly superior to traditional methods, offering a more authentic reproduction of the original timbre and providing a new effective approach for timbre restoration.

6. CONCLUSION

This paper has achieved innovative results in the field of ethnic musical instrument sound processing. Firstly, the proposed sound separation technique based on time-frequency analysis effectively extracts the target instrument's sound from mixed signals through precise analysis. This technique overcomes the confusion problems present in traditional sound separation methods, offering new possibilities for the extraction of individual instrument sounds. Secondly, the timbre synthesis method based on the *Transformer* model, with its powerful modeling capabilities of deep learning, successfully captures the subtle features of ethnic musical instrument timbres and generates highly realistic timbres, as verified through the synthesis effect analysis in experiments. Moreover, the paper introduces a novel continuity-based timbre restoration technique, significantly enhancing the naturalness and coherence during the restoration of damaged audio, with the cubic spline function demonstrating high fidelity to the original musical data in the restoration effect analysis.

While significant achievements have been made in this research, there is still room for further exploration in future studies. Firstly, time-frequency analysis and sound separation techniques can be further optimized, for instance, by utilizing more advanced signal processing algorithms and machine learning models to improve the precision of separation effects and the real-time nature of the separation process. Secondly, the *Transformer* model for timbre synthesis could explore new structures and training strategies to capture more complex timbre variations and enhance synthesis quality. Additionally, the application scope of timbre restoration techniques could be expanded from traditional ethnic instruments to a broader range of musical styles and instrument types, while considering the diverse damage scenarios that may occur in audio restoration, to develop more universal and robust restoration algorithms. Finally, integrating these technologies into music production and editing software for use by professionals and amateur music producers will be an important application direction for music technology research.

REFERENCES

[1] Zhou, S., Du, C. (2022). A virtual ethnic musical instrument platform based on web app. In 2022 7th International Conference on Multimedia Communication Technologies (ICMCT), Xiamen, China, pp. 10-14. https://doi.org/10.1109/ICMCT57031.2022.00011

[2] Setiawati, E., Karyaningsih, D., Ruiyat, S., Sutiawan, H., Sampurna, I., Susilawati, E.S. (2021). Introduction of traditional lebak musical instruments through aplikasi android for early childhood. In Journal of Physics: Conference Series, Tasikmalaya, Indonesia, p. 012082. https://doi.org/10.1088/1742-6596/1764/1/012082

[3] Zulfan, Baihaqi. (2019). Aceh Serune kale and Rapai ethnic musical instrument preservation method based on two-dimensional multimedia animation. Journal of Physics: Conference Series, Aceh, Indonesia, pp. 1-5. https://doi.org/10.1088/1742-6596/1232/1/012027

[4] Luo, J. (2022). Instrumental music dissemination of southwest ethnic minorities based on big data technology. In International Conference on Applications and Techniques in Cyber Intelligence, Fuyang, China, pp. 1011-1019. https://doi.org/10.1007/978-3-031-29097-8_121

[5] Kim, K., Park, M., Joung, H., Chae, Y., Hong, Y., Go, S., Lee, K. (2023). Show me the instruments: Musical instrument retrieval from mixture audio. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5. https://doi.org/10.1109/ICASSP49357.2023.10097162

[6] Sahoo, K.K., Hazra, R., Ijaz, M.F., Kim, S., Singh, P.K., Mahmud, M. (2022). MIC_FuzzyNET: Fuzzy integral based ensemble for automatic classification of musical instruments from audio signals. IEEE Access, 10: 100797-100811. https://doi.org/10.1109/ACCESS.2022.3208126

[7] Du, C., Guo, Y., Chen, X., Yu, K. (2023). Speaker adaptive text-to-speech with timbre-normalized vector-quantized feature. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 3446-3456. https://doi.org/10.1109/TASLP.2023.3308374

[8] Song, W., Yue, Y., Zhang, Y.J., Zhang, Z., Wu, Y., He, X. (2022). Multi-speaker multi-style speech synthesis with timbre and style disentanglement. In National Conference on Man-Machine Speech Communication, Hefei, China, pp. 132-140. https://doi.org/10.1007/978-981-99-2401-1_12

[9] Liu, J., Guo, Y., Chen, J., Wang, Z., Mao, A. (2022). Speech synthesis for speaker timbre translation across languages. In 2022 4th International Conference on Control and Robotics (ICCR), Guangzhou, China, pp. 320-324. https://doi.org/10.1109/ICCR55715.2022.10053890

[10] Liu, P., Choi, Y., Lee, J. (2023). Micro-variation of sound objects using component separation and diffusion models. In ICMC 2023: The Sound of Changes - International Computer Music Conference, Korea, pp. 199-203.

[11] Nawaz, R., Nisar, H., Voon, Y.V. (2018). The effect of music on human brain; Frequency domain and time series analysis using electroencephalogram. IEEE Access, 6: 45191-45205. https://doi.org/10.1109/ACCESS.2018.2855194

[12] Xue, C., Gu, Y., Gong, Z., Li, Z. (2023). Direction of arrival estimation of wideband hyperbolic frequency modulation signals using parameterized time-frequency analysis. Shengxue Xuebao/Acta Acustica, 48(1): 27-40. https://doi.org/10.15949/j.cnki.0371-0025.2023.01.004

[13] Gao, Y., Hu, Y., Wang, L., Huang, H., He, L. (2023). MTANet: Multi-band time-frequency attention network for singing melody extraction from polyphonic music. INTERSPEECH, Dublin, Ireland, pp. 5396-5400. https://doi.org/10.21437/Interspeech.2023-2494

[14] Kim, T.W., Kang, M.S., Lee, G.H. (2022). Adversarial

multi-task learning for disentangling timbre and pitch in singing voice synthesis. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3008-3012. https://doi.org/10.48550/arXiv.2206.11558

[15] Zhang, G., Qin, Y., Zhang, W., Wu, J., Li, M., Gai, Y., Jiang, F., Lee, T. (2023). iEmoTTS: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre. IEEE/ACM Transactions on Audio Speech and Language Processing, 31: 1693-1705. https://doi.org/10.1109/TASLP.2023.3268571

[16] Lee, J.Y., Bae, J.S., Mun, S., Lee, J., Lee, J.H., Cho, H.Y., Kim, C. (2023). Hierarchical timbre-cadence speaker encoder for zero-shot speech synthesis. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dublin, Ireland, pp. 4334-4338. https://doi.org/10.21437/Interspeech.2023-1128

[17] Tatar, K., Bisig, D., Pasquier, P. (2021). Latent timbre synthesis: Audio-based variational auto-encoders for music composition and sound design applications. Neural Computing and Applications, 33: 67-84. https://doi.org/10.1007/s00521-020-05424-2