Hindawi
Advances in Multimedia
Volume 2018, Article ID 5158164, 14 pages
https://doi.org/10.1155/2018/5158164

Research Article

Beijing Opera Synthesis Based on Straight Algorithm and Deep Learning

XueTing Wang[1], **Cong Jin**[2] **and Wei Zhao**[1]

[1]*College of Science and Technology, Communication University of China, Beijing, China*
[2]*Key Laboratory of Media Audio & Video, Communication University of China, Beijing, China*

Correspondence should be addressed to Cong Jin; jincong0623@cuc.edu.cn

Received 3 April 2018; Accepted 20 May 2018; Published 17 July 2018

Academic Editor: Yong Luo

Speech synthesis is an important research content in the field of human-computer interaction and has a wide range of applications. As one of its branches, singing synthesis plays an important role. Beijing Opera is a famous traditional Chinese opera, and it is called Chinese quintessence. The singing of Beijing Opera carries some features of speech but it has its own unique pronunciation rules and rhythms which differ from ordinary speech and singing. In this paper, we propose three models for the synthesis of Beijing Opera. Firstly, the speech signals of the source speaker and the target speaker are extracted by using the straight algorithm. And then through the training of GMM, we complete the voice control model to input the voice to be converted and output the voice after the voice conversion. Finally, by modeling the fundamental frequency, duration, and frequency separately, a melodic control model is constructed using GAN to realize the synthesis of the Beijing Opera fragment. We connect the fragments and superimpose the background music to achieve the synthesis of Beijing Opera. The experimental results show that the synthesized Beijing Opera has some audibility and can basically complete the composition of Beijing Opera. We also extend our models to human-AI cooperative music generation: given a target voice of human, we can generate a Beijing Opera which is sung by a new target voice.

1. Introduction

With the development of the times and the continuous innovation of science and technology, the demand for speech synthesis [1] is no longer simple to speak but can accomplish special voices such as singing and poetry. It is undoubtedly ingenious and novel to apply the method of singing synthesis [2] to Beijing Opera. Known as the quintessence of Chinese culture, Beijing Opera is one of the most famous traditional operas in China. And since its birth at the end of the 18th century, it has been favored by Chinese people and the people of other countries in East Asia. Beijing Opera has a long history and rich cultural connotation. In addition to the exquisite stage performances and vivid story plots, the music and singing of Beijing Opera are of great artistic value. In particular, it is a unique style of singing, which shows the extraordinary creativity of the Chinese nation, being the embodiment of the traditional artists' superb skills. It makes sense to use the straight algorithm, GMM, and GAN to synthesize Beijing Opera.

The synthesis of Beijing Opera can consist of three steps in Figure 1. First is voice conversion by using the straight algorithm, and then the synthesis of Beijing Opera fragments can be achieved through the tone control model and the melody control model. Finally, we connect the fragments and superimpose the background music to achieve the synthesis of Beijing Opera.

2. Synthesis of Beijing Opera with Straight Algorithm

2.1. Phoneme

2.1.1. Phoneme Profile. The phoneme is the smallest unit of speech or the smallest piece of speech that constitutes a

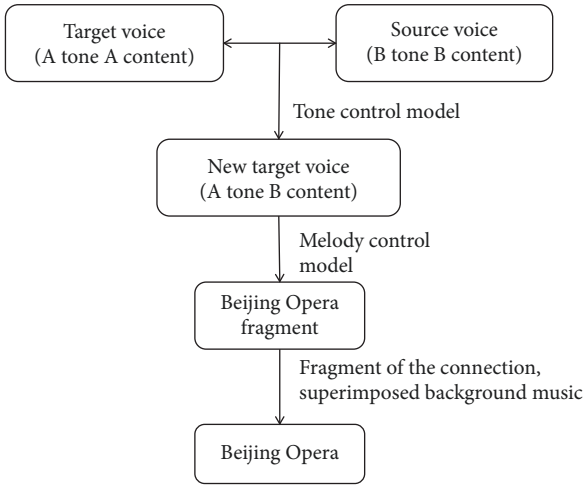

Figure 1: Beijing Opera synthesis.

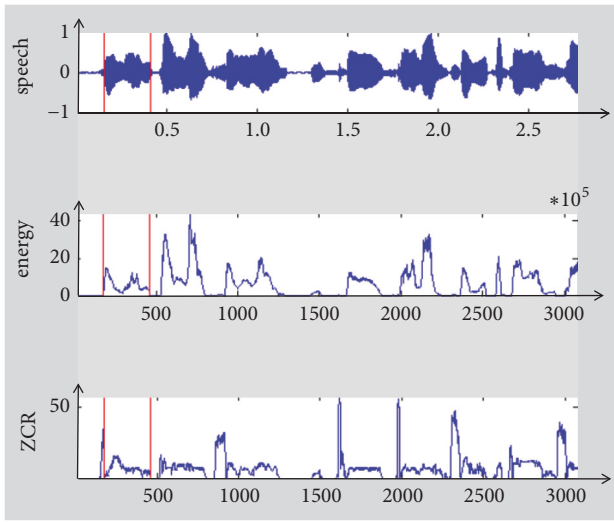

Figure 2: Time-domain waveforms, energy graphs, and zero-crossing rate graphs.

syllable and is the smallest linear speech unit that is divided from the perspective of sound quality. From the acoustic properties, phonemes are the smallest units of speech divided from the sound quality point of view. From a physiological point of view, a phonetic movement forms a phoneme. Phonemes are divided into vowels and consonants, two categories. Their classification is based on whether the airflow is obstructed by the various organs when the sound is emitted by humans. The unhindered factor is called the vowel and the obstructed one is called the consonant.

2.1.2. Phonemes Segmentation. Because the same phonemes have the same characteristics, and different factors and their combinations have different characteristics, we can divide each factor. The time-domain waveforms, energy graphs, and zero-crossing rate graphs of "jiao Zhang Sheng yin cang zai qi pan zhi xia" in "Matchmaker's Wall" sung by Beijing Opera were showed in Figure 2. From this, we can see that the consonant phonemes of the initial consonants are more irregular and the consonants formed by them have a periodic waveform. The former has the characteristics of large zero-crossing rate and low energy characteristics; the latter most of the energy is larger. In addition, if silence appears, both are small (red line is the beginning and end of a word).

2.2. Selection and Method of Characteristic Parameters

2.2.1. Choice of Personality Characteristics. Whether it is Beijing Opera or general voice, the speaker's personal habits and pronunciation styles are different on the one hand, and the speaker's position on the other (or the role of different actors in Beijing Opera) will result in each person having a handle on each phoneme a little difference. Generally speaking, the parameters that characterize the speaker's personality are the features of the syllabic, the suprasegmental, and the linguistic [3, 4].

Syllabic features: they describe the tonal characteristics of speech. The characteristic parameters mainly include the position of the formant, the bandwidth of the formant, the spectrum tilt, the pitch frequency, and the energy. Segment features are mainly related to the physiological and phonetic features of vocal organs and also to the speaker's emotional state. The features used in the tone control model in Section 3 are mainly for this reason.

Supersonic characteristics: they mainly refer to the way of speaking, such as the duration of phonemes, pitch, and stress, what people feel is the rate of speech, pitch, and volume changes. The features used in the melodic control model in Section 4 are mainly for this reason.

Language features: for example, idioms, dialects, accent, and so on.

However, Beijing Opera and voice are different in their purpose of pronunciation and expression. The pitch and pitch length of each word in Beijing Opera are controlled by the score in addition to its own pronunciation. The ordinary speech is mainly used to express the content of the speech, but Beijing Opera is more emotionally expressed by melody. Through the description of the above characteristics, the main considerations of this test sound quality mapping of the research factors are as follows.

Pitch: it is determined by the vibration frequency of the source in a period of time. The higher the vibration frequency, the higher the sound and the lower the converse. Beijing Opera's pitch and character roles, such as LaoSheng, are relatively low; Dan is relatively high.

Pitch length: the length of the sound is determined by the duration of the sound source vibration. The longer the duration, the longer the sound and the shorter the other hand. The average length of Beijing Opera per word is relatively long, and its variation range is relatively large.

Sound intensity: the strength of the sound depends on the vibration amplitude of the sound source; the greater the amplitude, the stronger the sound; on the other hand, the lower the amplitude, the smaller the sound. Since the amplitude of Beijing Opera is controlled by strong emotion, it is larger than the voice range. In general, the voice has only a relatively small amplitude range of uniform distribution.

Table 1: The correlations of subjective and objective amount of speech.

Objective amount	Subjective amount			
	pitch	volume	tone	duration
fundamental frequency	+++	+	++	+
amplitude	+	+++	+	+
spectral envelope	++	+	+++	+
time	+	+	+	+++
Relevance is positively related to the number of '+' s				

Tone: the frequency performance of different sounds always has distinctive characteristics in waveforms. For example, different Beijing Opera characters sing the same passage according to the difference between the two timbres.

By combining the subjective amount of speech with the objective amount we have analyzed, the correlations can be obtained in Table 1.

Acoustic characteristics of speech signal are an indispensable research object for speech analysis and speech transformation. It mainly displays prosody and spectrum. Prosody perceives performance as pitch, duration, and volume. Acoustically, the rhythm corresponds to the fundamental frequency, duration, and amplitude. The spectral envelope is perceived as a tonal characteristic.

2.2.2. MFCC Feature Extraction. MFCC is an acronym for Mel Frequency Cepstrum Coefficient (MFCC), which is based on human auditory properties and is nonlinearly related to Hz frequency. The Mel Frequency Cepstral Coefficients (MFCCs) use this relationship between them to calculate the resulting Hz spectral signature. Its extraction principle is as follows.

(1) Pre-Emphasis. Pre-emphasis processing is to pass the speech signal through a high-pass filter as

$$H(z) = 1 - \mu z^{-1} \tag{1}$$

The value of μ is between 0.9 and 1.0; we usually take 0.97. The purpose of pre-emphasis is to raise the high-frequency part, flatten the spectrum of the signal, and keep it in the whole low-to-high frequency band, with the same signal-to-noise ratio. At the same time, it is also to eliminate the vocal cords and lips in the process of occurrence, to compensate for the voice signal suppressed by the high frequency part of the system, but also to highlight the high-frequency formant.

(2) Framing. The first N sampling points set into a unit of observation, known as the frame. Under normal circumstances, the value of N is 256 or 512, covering about 20 ~ 30ms or so. In order to avoid the change of two adjacent frames being too large, there is an overlapping area between two adjacent frames. The overlapping area contains M sampling points, and the value of M is usually about 1/2 or 1/3 of N. Usually speech recognition [5] voice signal sampling frequency is 8KHz or 16KHz. For 8KHz, if the frame length is 256 samples, the corresponding time length is (256/8000)∗1000=32ms.

(3) Windowing (Hamming Window). Multiply each frame by a Hamming window to increase continuity at the left and right ends of the frame. Assuming the framed signal is S (n), n = 0,1, ..., N-1, N is the size of the frame then multiplied by the Hamming window is $S'(n) = S(n) \times W(n)$. The form of W (n) is as

$$W(n, a) = (1 - a) - a \times \cos\left[\frac{2\pi n}{N-1}\right], \quad 0 \le n \le N-1 \tag{2}$$

Different 'a' value will produce a different Hamming window. In general, 'a' takes 0.46

$$s'_n = \left\{0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right)\right\} * s_n \tag{3}$$

(4) Fast Fourier Transform. As the signal in the time-domain transformation is usually difficult to see the characteristics of the signal, so it is usually converted to the frequency domain to observe the energy distribution; different energy distribution can represent different voice characteristics. Therefore, after multiplying the Hamming window, each frame must also be subjected to fast Fourier transform to obtain the spectral energy distribution. The signal of each frame after windowing is subjected to fast Fourier transform to obtain the spectrum of each frame. And the spectrum of the speech signal is modeled to obtain the power spectrum of the speech signal. Set the DFT of the voice signal as

$$X_a(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi k/N}, \quad 0 \le k \le N \tag{4}$$

where x (n) is the input speech signal and N is the number of points in the Fourier transform.

(5) Triangular Bandpass Filter. The energy spectrum is passed through a set of Mel-scale triangular filter banks to define a filter bank with M filters (the number of filters is similar to the number of critical bands). The filter used is a triangular filter. M usually takes 22-26. The spacing between each f (m) decreases with decreasing 'm', broadening as m increases, as shown in Figure 3.

The frequency response of the triangular filter is defined as

$$H_{\mathrm{m}}(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{2(k - f(m-1))}{[f(m+1) - f(m-1)][f(m) - f(m-1)]}, & f(m-1) \le k \le f(m) \\ \dfrac{2(f(m+1) - k)}{[f(m+1) - f(m-1)][f(m) - f(m-1)]}, & f(m) \le k \le f(m+1) \\ 0, & k \ge f(m+1) \end{cases} \tag{5}$$

(6) Calculate logarithmic energy output from each filter bank as

$$s(m) = \ln\left[\sum_{k=0}^{N-1} \left|X_a(k)^2\right| H_m(k)\right], \quad 0 \le m \le M \tag{6}$$

(7) The MFCC coefficients are obtained by discrete cosine transform (DCT) as

$$C(n) = \sum_{m=0}^{N-1} s(m) \cos\left[\frac{\pi n(m - 0.5)}{M}\right], \quad n = 1, 2, \ldots, L \tag{7}$$

The above logarithmic energy is taken into DCT to obtain the L-order Mel-scale Cepstrum parameter. The L-order means the MFCC coefficient order, usually 12-16. Here M is the number of triangular filters.

(8) Logarithmic Energy. In addition, the volume (i.e., energy) of a frame is also an important feature of speech and is very easy to calculate. Therefore, the logarithmic energy of one frame is usually added so that the basic speech features of each frame have one more dimension, including one logarithmic energy and the remaining cepstrum parameters.

(9) Dynamic Segmentation Parameters Extraction (including First-Order Difference and Second-Order Difference). The standard cepstrum parameter MFCC only reflects the static characteristics of the speech parameters. The dynamic characteristics of the speech can be described by the difference spectrum of these static characteristics. Experiments show that combining dynamic and static features can effectively improve the system's recognition performance. The calculation of the difference parameter can use the following formula as

$$\mathrm{d_t} = \begin{cases} C_{t+1} - C_t, & t < K \\ \dfrac{\sum_{k=1}^{K} k(C_{t+k} - C_{t-k})}{2\sum_{k=1}^{K} k^2}, & \text{others} \\ C_t - C_{t-1}, & t \ge Q - k \end{cases} \tag{8}$$

where dt is the t-th first-order difference; Ct is the t-th cepstrum coefficient; Q is the order of the cepstral coefficients; and K is the time difference of the first derivative, which can be 1 or 2. Substituting the result in the above equation yields the second-order difference parameter.

2.3. Signal Characteristics Analysis. According to the previous research on speech signal processing technology, people mainly focus on the signal analysis in the time domain and frequency domain of these two methods.

2.3.1. Time-Domain Analysis. In the time domain, the horizontal axis is the time and the vertical axis is the amplitude. By observing the waveform in the time domain, we can obtain some important features of the speech signal, such as the duration, the starting and ending positions of the syllables, the sound intensity (energy), and vowels (see Figure 4).

2.3.2. Frequency Domain Analysis. The voice signal spectrum, power spectrum, cepstrum, spectral envelope, and so on are included. It is generally considered that the frequency spectrum of the speech signal is the product of the frequency response of the channel system and the spectrum of the excitation source, while the frequency response of the channel system and the excitation source are time-varying. Therefore, frequency domain analysis of speech signals is often performed using short-time Fourier transform (STFT). It is defined as

$$X_{\mathrm{n}}\left(e^{j\omega}\right) = \sum_{m=-\infty}^{+\infty} x(m)\, w(n - m)\, e^{-j\omega m} \tag{9}$$

The study of Chinese song synthesis algorithm is based on parameter modification, where we can see that short-time Fourier transform has two independent variables (n and w), so it is both a discrete function about time n and a continuous function about angular frequency. In the formula, w (n) is a window function, and n takes different values and removes different voice short segments, where the subscript n is different from the standard Fourier transform. Since the shape of the window has an influence on the short-time spectrum, the window function should have the following characteristics:

(1) High frequency resolution; the main lobe is narrow and sharp.

(2) Side lobe attenuation is large, and spectrum leakage caused by other frequency components is small. These two conditions are in fact contradictory to each other and cannot be satisfied at the same time. Therefore, we often adopt a compromise approach and often choose a Hamming window.

However, both time-domain analysis and frequency domain analysis have their own limitations: time-domain analysis does not have an intuitive visualization of the

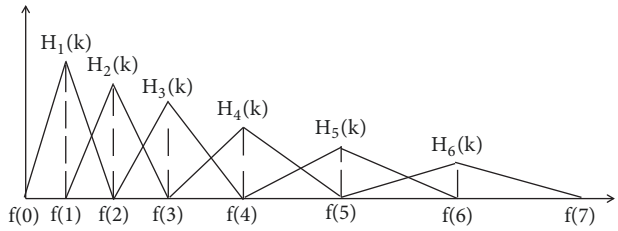

Figure 3: Mel frequency filter bank.

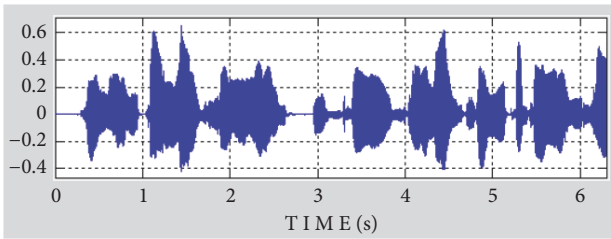

Figure 4: "Jiao Zhang Sheng yin cang zai qi pan zhi xia" time-domain diagram.

frequency characteristics of speech signals; and frequency domain analysis lacks the variation of speech signals over time. As a result, the experiment of the Beijing Opera synthesis analyzed the speech signal using the later improved method of analyzing the spectrum.

2.3.3. Spectrum Analysis. The Fourier analysis display of the speech signal is called a sonogram or spectrogram. A spectrogram is a three-dimensional spectrum that represents a graph of the frequency spectrum of a voice over time, with the vertical axis as the frequency and the horizontal axis as the time. The intensity of any given frequency component at a given moment is expressed in terms of the grayness or hue of the corresponding point. The spectrum shows a great deal of information related to the characteristics of the speech sentence. It combines the characteristics of spectrograms and time-domain waveforms to clearly show how the speech spectrum changes over time or is a dynamic spectrum. From the spectrum we can get: formant, fundamental frequency, and other parameters in Figure 5.

2.4. Straight Algorithm Introduction. Straight is an acronym for "Speech Transformation and Representation based on Adaptive Interpolation of weighted spectrogram." It is a more accurate method of speech analysis and synthesis proposed by Japanese scholar Kawara Eiji in 1997. The straight algorithm builds on the source/filter model. Among them, the source comes from the vocal cords vibration, and the filter refers to the channel transfer function. It can adaptively interpolate and smooth the speech short-duration spectrum in the time domain and the frequency domain so as to extract the spectral envelope more accurately and adjust the speech duration, fundamental frequency, and spectral parameters to a great extent without affecting the quality of the synthesized speech. The straight analysis synthesis algorithm consists of three steps: fundamental frequency extraction, spectral

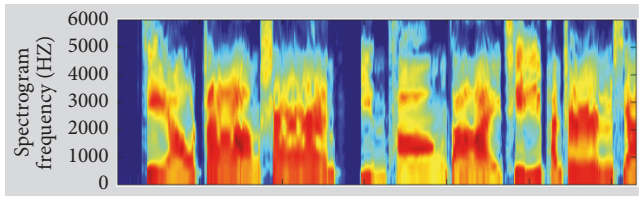

Figure 5: "Jiao Zhang Sheng yin cang zai qi pan zhi xia" spectrogram.

parameter estimation, and speech synthesis. The first two of them are described in detail below, and only the synthesis process will be described in Figure 6.

First of all, the speech signal is input, the speech fundamental frequency F0 and spectral envelope are extracted by straight algorithm, and the parameters are modulated to generate a new sound source and time-varying filter. According to the original filter model, we use (10) to synth voice:

$$y(t) = \sum_{t_i \in Q} \frac{1}{\sqrt{G(f_D(t_i))}} v_{t_i}(t - T(t_i)) \tag{10}$$

$v_{t_i}, T(t_i)$ is shown as

$$v_{t_i}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} V(\omega, t_i)\, \varphi(\omega)\, e^{j\omega(t)} d\omega \tag{11}$$

$$T(t_i) = \sum_{t_k \in Q, k<i} \frac{1}{\sqrt{G(f_0(t_k))}} \tag{12}$$

In the formula, Q represents the position of a group of samples in the synthesis excitation, and G represents the pitch modulation. The F0 after modulation can be matched with any F0 of the original language arbitrarily. All-pass filter is used to control the time structure of fine pitch and original signal, such as a frequency-proportional linear phase shift, used to control the fine structure of F0. $V(\omega, t_i)$ is the corresponding Fourier transform of the minimum phase pulse; as in (12) $A[S(u(\omega), r(t)), u(\omega), r(t)]$ is calculated from the modulation amplitude spectrum, where A, u, and r represent the modulation of amplitude, frequency, and time, respectively, as (13), (14), and (15).

$$V(\omega, t) = e^{(1/\sqrt{2\pi}) \int_0^{\infty} h_t(q) e^{j\omega q} dq} \tag{13}$$

$$h_t(q) = \begin{cases} 0, & (q < 0) \\ c_t(0), & (q = 0) \\ 2c_t(q), & (q > 0) \end{cases} \tag{14}$$

$$c_t(q) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} e^{-j\omega q} \lg A\{S[u(\omega), r(t)], u(\omega), r(t)\} d\omega \tag{15}$$

q is the frequency. Straight audiometry experiments show that even in the case of high-sensitivity headphones, the

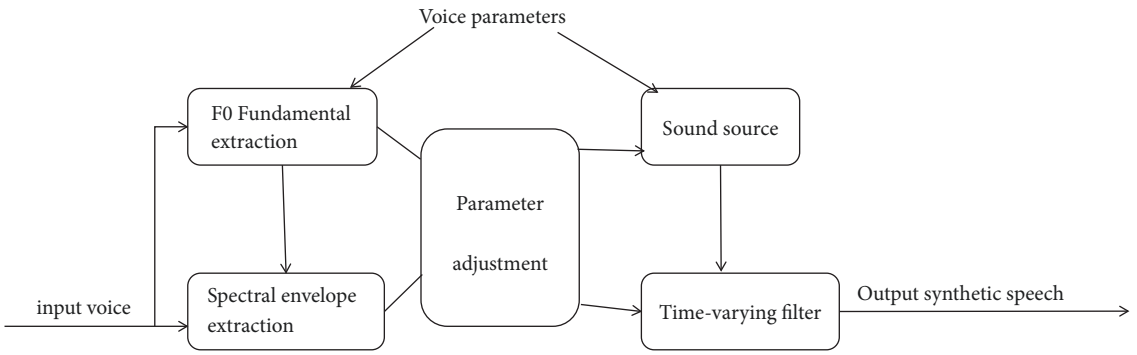

Figure 6: Straight synthesis system.

synthesized speech signal is almost indistinguishable from the original signal.

3. Tone Control Model

Voice tonal conversion refers to the voice signal processing technology to deal with the voice, to maintain the same semantic content, but only change the tone, so that a person's voice signal (source voice) after the sound conversion processing sounds like another person voice (target voice). This chapter introduces the extraction of the parameters that are closely related to the timbre by using the straight algorithm, and then the training model of the extracted parameters by using GMM to get the corresponding relationship between the source voice and the target voice. Finally, the new parameters are straight synthesis, in order to achieve voice conversion. It can be seen from Section 2 that the tone characteristics in speech mainly correspond to the parameters "fundamental F0" and "channel spectrum".

3.1. The Fundamental Frequency and Channel Spectrum Extraction

3.1.1. Extraction of the Fundamental Frequency. Straight algorithm has a good time-domain resolution and fundamental frequency trajectory, which is based on wavelet transform to analyze, first found from the extracted audio frequency to find the base frequency, and then calculated the instantaneous frequency, as the fundamental frequency.

Fundamentals of the extraction can be divided into three parts: F0 coarse positioning, F0 track smooth, and F0 fine positioning. The coarse positioning of F0 refers to the wavelet transform of the voice signal to obtain the wavelet coefficients; then, the wavelet coefficients are transformed into a set of instantaneous frequencies for selecting F0 for each frame. F0 trajectory smoothing is based on the calculated high-frequency energy ratio, the minimum noise energy equivalent, in the instantaneous frequency selected the most likely F0, and thus constitutes a smooth pitch trajectory. F0 fine positioning through FFT transforms the current F0 fine-tuning. The process is as follows.

Input signal is s(t), the output composite signal is $D(t, \tau_c)$, where gAG(t) is an analysis of the wavelet which is gotten by the input signal through Gabor filter, and τ_c is analysis cycle of the analyzed wavelet as

$$D(t, \tau_c) = |\tau_c|^{-1/2} \int_{-\infty}^{+\infty} s(t)\, gAG \frac{t-\mu}{\tau_c} d\mu \tag{16}$$

gAG(t) is (17) and shown as (18):

$$g_{AG}(t) = g\left(t - \frac{1}{4}\right) - g\left(t + \frac{1}{4}\right) \tag{17}$$

$$g(t) = e^{-\pi(t/\eta)^2} e^{-j2\pi t} \tag{18}$$

Among them, η is the frequency resolution of the Gabor filter, which is usually larger than 1 according to the characteristics of the filter.

Through calculation, the variable "fundamentalness" is introduced and denoted by $M(t, \tau_0)$ as

$$\begin{aligned} M = &- \log\left[\int_\Omega \left(\frac{d|D|}{du}\right) du\right] + \log\left[\int_\Omega |D|^2\, du\right] \\ &- \log\left[\int_\Omega \left(\frac{d \arg |D|}{du}\right)^2 du\right] + 2\log\tau_0 \\ &+ \log\Omega(\tau_0) \end{aligned} \tag{19}$$

The first term is the amplitude modulation (AM) value; the second term is the total energy, used to normalize the value of AM; the third term is the frequency modulation (FM) value; the fourth term is the square of the fundamental frequency, used to normalize the value of FM; the fifth is the normalization factor of the time-domain integration interval. By the formula the following can be drawn: when AM, FM take the minimum, M takes the maximum, namely, getting the fundamental part.

However, in practice, F0 always changes rapidly, so in order to reduce the impact on M, the formula makes some adjustments as (20), (21), and (22):

$$\begin{aligned} M = & -\log\left[\int_{\Omega}\left(\frac{d|D|}{du}-\mu_{FM}\right)^{2}du\right] \\ & +\log\left[\int_{\Omega}|D|^{2}du\right] \\ & -\log\left[\int_{\Omega}\left(\frac{d\arg|D|}{du}-\mu_{FM}\right)^{2}du\right] \\ & +2\log\tau_{0}+\log\Omega\left(\tau_{0}\right) \end{aligned} \tag{20}$$

$$\mu_{AM}=\frac{1}{\Omega}\int_{\Omega}\left(\frac{d|D|}{du}\right) \tag{21}$$

$$\mu_{FM}=\frac{1}{\Omega}\int_{\Omega}\left(\frac{d^{2}\arg(\mathrm{D})}{du^{2}}\right) \tag{22}$$

Finally use τ_0 to calculate the instantaneous frequency $\omega(t)$, and get the fundamental frequency F0 by (23), (24), and (25):

$$f_{0}=\frac{\omega_{0}(t)}{2\pi} \tag{23}$$

$$\omega(t)=2f_{s}\arcsin\frac{\left|y_{d}(t)\right|}{2} \tag{24}$$

$$\mathrm{y}_{d}(t)=\frac{D\left(t+\Delta t/2,\tau_{0}\right)}{\left|D\left(t+\Delta t/2,\tau_{0}\right)\right|}-\frac{D\left(t-\Delta t/2,\tau_{0}\right)}{\left|D\left(t-\Delta t/2,\tau_{0}\right)\right|} \tag{25}$$

3.1.2. Channel Spectral Parameter Extraction. The voice of the sound source information and channel spectrum information extracted and then make adjustments to achieve voice adjustment, which is the previous method. However, since the two are often highly correlated, they cannot be independently modified, thus affecting the final result.

The relationship among the voice signal s(t), the channel parameter v(t), and the sound source parameter p(t) is as

$$s(t)=p(t)*v(t) \tag{26}$$

Since it is difficult to find v(t) directly, the straight algorithm calculates the frequency domain expression of v(t) by short-time spectral analysis of s(t). The method to calculate the short-term spectrum is (27) and (28):

$$sw\left(t,t'\right)=s(t)\,w\left(t,t'\right) \tag{27}$$

$$SW\left(\omega,t'\right)=FFT\left[sw\left(t,t'\right)\right]=S\left(\omega,t'\right)W\left(\omega,t'\right) \tag{28}$$

The short-term spectrum shows the periodicity related to the fundamental frequency in the time domain and the frequency domain, respectively. The short-time spectrum window function used is (29) and (30):

$$w(t)=\frac{1}{f_{0}}e^{-\pi(t/f_{0})^{2}} \tag{29}$$

$$W(\omega)=\frac{f_{0}}{\sqrt{2\pi}}e^{-\pi(\omega/\omega_{0})^{2}} \tag{30}$$

However, since both the channel spectrum and the sound source spectrum are related to the fundamental frequency at this time, it cannot be considered that they have been separated. Instead, they need to be further cyclically removed in the time domain and the frequency domain to achieve the separation.

Periodic removal of the time domain requires the design of pitch-sync smoothing windows and compensation windows, respectively, as (31), (32), and (33):

$$w_{p}(t)=e^{-\pi(t/\tau_{0})}*h\left(\frac{t}{\tau_{0}}\right) \tag{31}$$

$$h(t)=\begin{cases}1-|t| & (t<1)\\ 0 & (otherwise)\end{cases} \tag{32}$$

$$w_{\mathrm{c}}(t)=w_{p}(t)\sin\left(\pi\times\frac{t}{\tau_{0}}\right) \tag{33}$$

Then the short-time amplitude spectrum $|S_{\mathrm{p}}(\omega,t')|$ and $|S_{\mathrm{p}}(\omega,t')|$, respectively, is obtained by the two windows and finally we get the short-term amplitude spectrum with the periodicity removed as

$$\left|S_{\mathrm{r}}\left(\omega,t'\right)\right|=\sqrt{\left|S_{\mathrm{r}}\left(\omega,t'\right)\right|^{2}+\xi\left|S_{\mathrm{r}}\left(\omega,t'\right)\right|^{2}} \tag{34}$$

Among them, ξ is the mixing factor; when ξ is taking 0.13655, there is the optimal solution.

Similarly, the frequency domain also needs smoothing windows $V(\omega)$ and compensation windows $U(\omega)$ to remove the periodicity in the short-time spectral $SW(\omega)$ domain and finally remove the periodic spectral envelope $SS'(\omega)$ as

$$SS'(\omega)=SW(\omega)*V(\omega)*U(\omega) \tag{35}$$

Finally, the logarithmic amplitude compression and distortion frequency discrete cosine transform the channel spectral parameters into MFCC parameters for the subsequent use.(MFCC is described in detail in Section 2).

3.2. GMM Achieve Parameter Conversion

3.2.1. GMM Profile. The Gaussian Mixture Model (GMM) [6] can be expressed as a linear combination of different Gaussian probability functions in

$$P\left(\frac{X}{\lambda}\right)=\sum_{i=1}^{M}\omega_{i}b_{i}(X) \tag{36}$$

where X is a random vector of n dimensions, ω_i is a mixture weight, $\sum_{i=1}^{M} \omega_i = 1$, $b_i(X)$ is a subdistribution of GMM, and each subdistribution is a Gaussian distribution as

$$b_i(X) = \frac{1}{(2\pi)^{n/2} \left|\sum_i\right|^{1/2}} e^{-(1/2)(X-\mu_i)^T \sum_i^{-1} (X-\mu_i)} \quad (37)$$

where μ_i is the mean vector and $\sum_i$ is the covariance matrix.

Although the types of phonemes are definite, each phoneme varies in different situations due to the context. We use GMM to construct the acoustic characteristics of the speaker to find the most likely mapping at each time.

3.2.2. Conversion Function to Establish. GMM refers to the estimation of the probability density distribution of the sample, and the estimated model (training model) is the weighted sum of several Gaussian models. It maps matrices of source speech and target speech, thereby increasing the accuracy and robustness of the algorithm and completing the connection between the two phonetics.

(1) The Conversion of Fundamental Frequency. Here the single Gaussian model method is used to convert the fundamental frequency, and the converted fundamental frequency is obtained through the mean and variance of the target person $(\mu_{tgt}, \sigma_{tgt})$ and the speaker $(\mu_{src}, \sigma_{src})$ is (38):

$$f_{0,conv}(t) = \sqrt{\frac{\sigma_{tgt}^2}{\sigma_{src}^2}} \times f_{0,src}(t) + \mu_{tgt} - \sqrt{\frac{\sigma_{tgt}^2}{\sigma_{src}^2}} \times \mu_{src} \quad (38)$$

(2) Channel Spectrum Conversion. The model's mapping rule is a linear regression function; the purpose is to predict the required output data by inputting data. The spectrum conversion function is defined as

$$\begin{aligned} F(X) &= E\left\{\frac{Y}{X}\right\} = \int Y * P\left(\frac{Y}{X}\right) dY \\ &= \sum_{i=1}^{M} P_i(X) \left[\mu_i^Y + \sum_i^{YX} \left(\sum_i^{XX}\right)^{-1} \left(X - \mu_i^X\right)\right] \end{aligned} \quad (39)$$

$$P_i(X) = \frac{\omega_i b_i(X_t)}{\sum_{k=1}^{M} \omega_k b_k(X_t)} \quad (40)$$

$$\mu_i = \begin{bmatrix} \mu_i^X \\ \mu_i^Y \end{bmatrix},$$

$$\sum_i = \begin{bmatrix} \sum_i^{XX} & \sum_i^{XY} \\ \sum_i^{YX} & \sum_i^{YY} \end{bmatrix}, \quad i = 1, \ldots, M \quad (41)$$

μ_i^X and μ_i^Y are the mean of the i-th Gaussian component of the source speaker and the target speaker; $\sum_i^{XY}$ is the variance matrix of the i-th Gaussian component of the source speaker; $\sum_i^{XY}$ is the covariance matrix of the ith Gaussian component of the source speaker and the target speaker covariance matrix; $P_i(X)$ is the feature vector probability of X belonging to the i-th Gaussian components of the GMM.

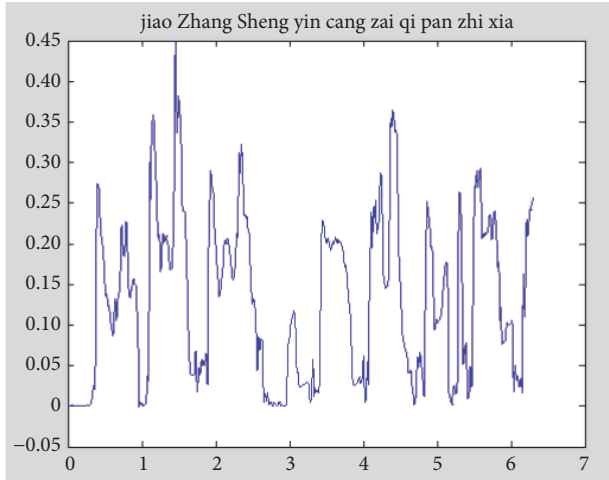

Figure 7: "Jiao Zhang Sheng yin cang zai qi pan zhi xia" envelope.

4. Melody Control Model

The composition of Beijing Opera has similarities with the synthesis of general singing voice [7, 8]. That is, through the superimposition of voice and melody, the new pitch of each word is reconstructed. Through the analysis of the second chapter, it is found that the major factors affecting the melody are the fundamental frequency, duration, and energy. Among them, the fundamental frequency of melody has the greatest impact; it can indicate the frequency of human vocal vibration and duration and pronunciation of each word of each word length; you can control the rhythm of Beijing Opera, which represents the speed of human voice. Energy and sound intensity were positively correlated, representing the emotions.

4.1. The Fundamental Frequency Conversion Model. Although both speech and Beijing Opera are issued through the same human organs, the speech pays more attention to the prose, while the Beijing Opera emphasizes the emotional expression of the melody. Most of the features in the melody are in the fundamental frequency. The fundamental envelope of a Beijing Opera corresponds to the melody, which includes tone, pitch, and tremolo [9]. However, the pitch of a note in a note is a constant, and their comparison is as in Figure 7.

From this we can see that we can use the fundamental frequency to control the melody of a Beijing Opera, but the acoustic effects such as vibrato need to be considered. Therefore, the control design of the fundamental frequency [10] is as in Figure 8.

4.2. Time Control Model. Each word in Chinese usually has different syllables, and the initials and vowels in each syllable

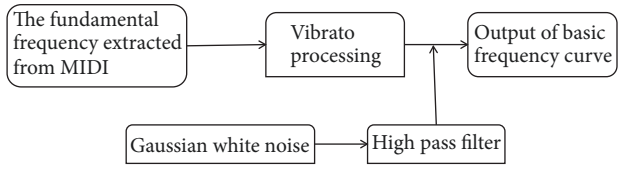

Figure 8: The control design of the fundamental frequency.

Table 2: Duration parameters.

Duration parameters	
Before modification	After modification
dur_a	k∗dur_a
dur_b	dur_b
dur_c	dur_t – (k∗dur_a) – dur_b

dur_a: initial part duration, dur_b: initial part to vowel transition part duration, dur_c: final part duration, and dur_t: target total duration.

also play different roles. The initials, whether normal or Beijing Opera, usually play a supporting role, while vowels carry pitch and most of the pitch information. In order to ensure the naturalness of Beijing Opera, we use the note duration to control the length of each word and make the rules for the vowel length shown in Table 2.

Initial part of the length of time is in accordance with the proportion [11] (k in the table) to be modified, k is a lot of voice, and song comparison experiments are obtained. The duration of the area with the initials to vowels transition remains unchanged. The length of the vowel section varies so that the total duration of the syllable can correspond to the duration of each note in the score.

The method of dividing the vowel boundaries is introduced in Section 2 and will not be repeated here.

4.3. Spectrum Control Model. The vocal tract is a resonant cavity and the spectral envelope reflects its resonant properties. Studies have found good vibes singing, the spectrum in the vicinity of 2.5-3 kHz has a special resonance farm, and singing spectrum changes will directly affect the people of the party's results. In order to synthesize music of high naturalness, the spectral envelope of the speech signal is usually corrected according to the unique spectral characteristics of the singing voice.

4.4. GAN Model

4.4.1. Introduction of GAN Network. Generative adversarial networks, abbreviated as GAN [12–17], are currently a hot research direction in artificial intelligence. The GAN consists of generators and discriminators. The training process is inputting random noise, obtaining pseudo data by the generator, taking a part of the real data from the true data, mixing the two and sending the data to the discriminator, giving a true or false determination result, and, according to this result, the return loss. The purpose of GAN is to estimate the potential distribution of data samples and generate new data samples. It is being extensively studied in the fields of image and visual computing, speech, and language processing and has a huge application prospect. This study uses GAN to synthesize music to compose Beijing Opera music.

4.4.2. Selection of Test Datasets. The Beijing Opera score dataset that needs to be used in this study is the recording and collection of 5,000 Beijing Opera background music tracks. The dataset is processed as shown in the Figure 9: dataset preparation and data preprocessing

First of all, because sometimes some instruments have only a few notes in a piece of music, this situation makes the data too sparse and affects the training process. Therefore, it is necessary to solve this data imbalance problem by merging the sound tracks of similar instruments. Each of the multitrack Beijing Opera scores is incorporated into five musical instruments: huqins, flutes, suonas, drums, and cymbals. These five types of instruments are the most commonly used musical instruments in Beijing Opera music.

Then, we will filter the datasets after the merged tracks to select the music with the best matching confidence. In addition, because the Beijing Opera arias need to be synthesized, the scores in the part of the Beijing Opera without lyrics are not what we need. Also select the soundtracks of Beijing Opera lyrics.

Finally, in order to obtain a meaningful music segment to train the time model, it is necessary to divide the Peking Opera score and obtain the corresponding music segment. Think of the 4 bars as a passage and cut the longer passage into the appropriate length. Because pitches that are too high or too low are not common and are therefore less than C1 or higher than C8, the target output tensor is 4 (bar) × 96 (time step) × 84 (pitch) × 5 (track). This completes the preparation and preprocessing of the dataset.

4.4.3. Training and Testing of GAN Structure and Datasets. The GAN structure diagram used in this study is as in Figure 10.

The basic framework of the GAN includes a pair of models: a generative model and a discriminative model. The main purpose is to generate pseudo data consistent with the true data distribution by the discriminator D auxiliary generator G. The input of the model is a random Gaussian white noise signal z; the noise signal is mapped to a new data space via the generator G to generate the generated data G(z). Next, a discriminator D outputs a probability value based on the input of the true data x and the generated data G(z), respectively, indicating that the D judges whether the input is real data or the confidence of generating false data. In this way, it is judged whether the performance of the G-generated data is good or bad. When the final D cannot distinguish between the real data x and the generated data G(z), the generator G is considered to be optimal.

The goal of D is to distinguish between real data and false data so that D(x) is as large as possible, while D(G(z)) is as small as possible, and the difference between the two is as large as possible, whereas G's goal is to make the data it produces in D. The goal of G is to make the performance 'D(G(z))' of its own data on D consistent with

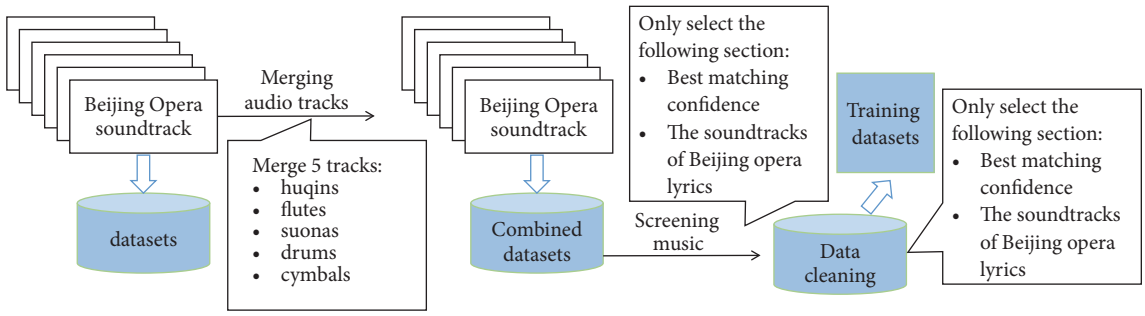

Figure 9: Illustration of the dataset preparation and data preprocessing procedure.

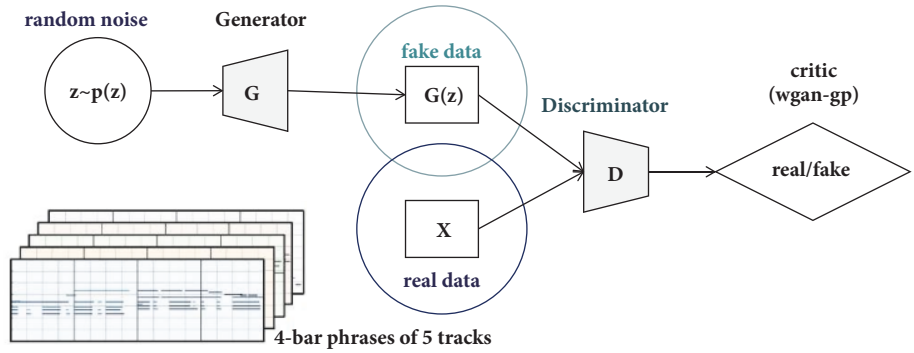

Figure 10: GAN structure diagram.

the performance 'D(x)' of the real data, so that D cannot distinguish between generated data and real data. Therefore, the optimization process of the module is a process of mutual competition and confrontation. The performance of G and D is continuously improved during repeated iteration until the final D(G(z)) is consistent with the performance D(x) of the real data. And both G and D cannot be further optimized.

The training process can be modeled as a simple MinMax problem in

$$\min_{G} \max_{D} D(x) - D(G(z)) \tag{42}$$

The MinMax optimization formula is defined as follows:

$$\min_{qG} \max_{qD} V(D,G) = \min_{G} \max_{D} \left\{ E_{x:pG} \left[\log D(x) \right] + E_{x:pG} \left[\log \left(1 - D(G(z)) \right) \right] \right\} \tag{43}$$

The GAN does not require a pre-set data distribution; that is, it does not need to formulate a description of p(x) but directly adopts it. Theoretically, it can completely approximate real data. This is the biggest advantage of the GAN.

The training and testing process of the GAN generated music dataset is as in Figure 11.

The generator-generated chord section data and specific music style data, generator-generated multiple track chord section data, and multiple tracks of music groove data are sent to the GAN for training. Reach music that generates specific styles and corresponding grooves.

5. Experiment

5.1. Tone Control Model

5.1.1. Experimental Process. The voice library used in the experiment simulation of this article is recorded by the former in the environment of the entire anechoic room, and comprehensive consideration of the previous factors can better meet the actual needs of the speech conversion system. The voice library is recorded by a woman in standard Mandarin accent and contains numbers, professional nouns, everyday words, etc., as source speech. Then find another person to record a small number of statements as the voice to be converted, and Figure 12 is the tone conversion process.

5.1.2. Experimental Results. Figures 11, 12, and 13 speech of the source speaker, respectively, and the target speaker is based on the speech spectrogram STRAIGHT and convert speech GMM model obtained. All voices are sampled at 16khz and quantized with 16 bits. Set the voice to 5s during the experiment. Their MFCCs are in Figures 13, 14, and 15.

They show the MFCC three-dimensional map of the source speech, the target speech, and the converted speech. The horizontal axis represents the audio duration, the vertical axis represents the frequency, and the color represents the corresponding energy. From the comparison of the graphs, it can be directly seen that the vocalogram shape of the converted MFCC parameters is closer to the target speech, indicating that the converted speech features tend to be the target speech features.

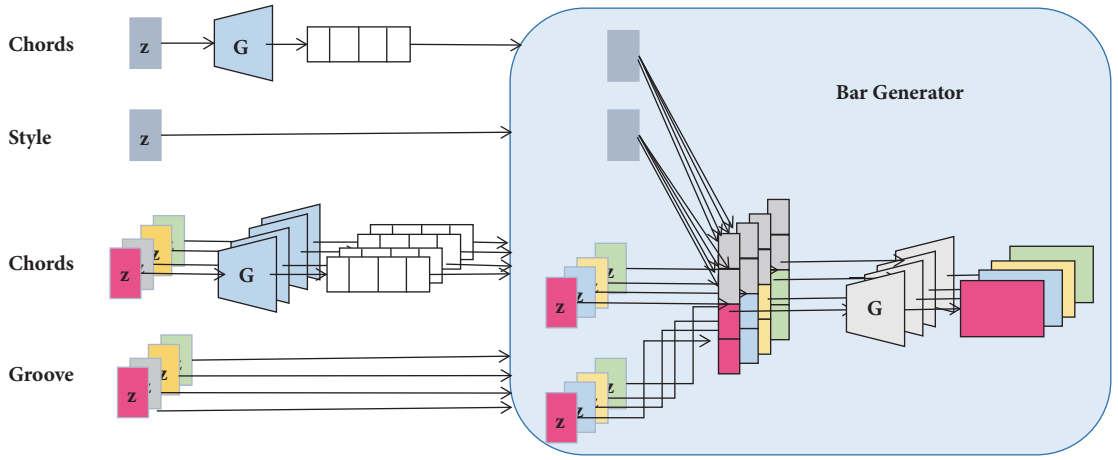

Figure 11: Raining and testing process of the GAN.

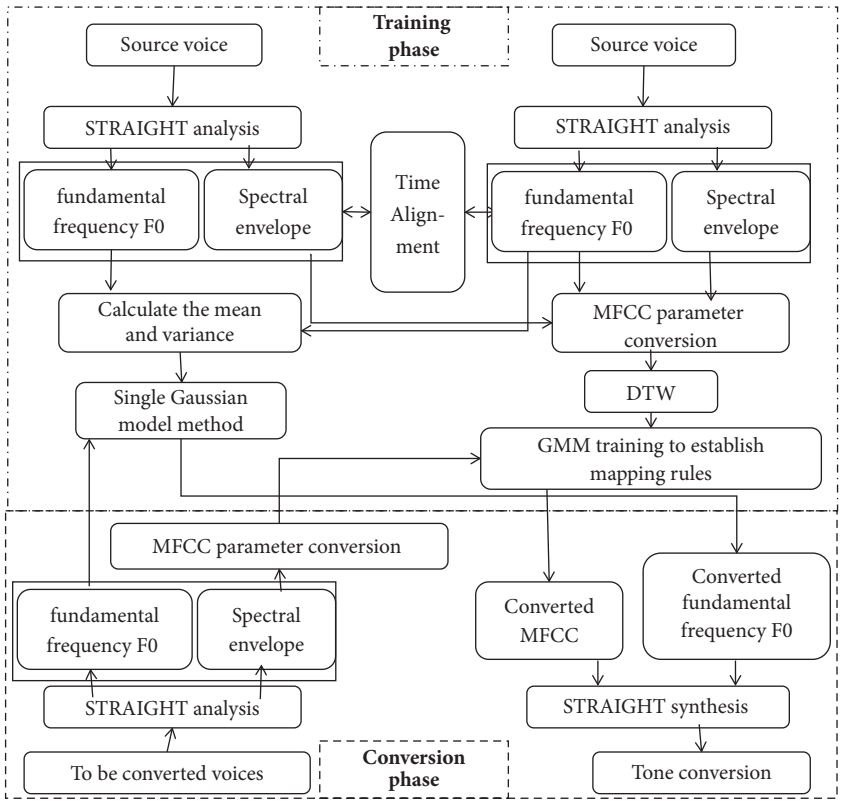

Figure 12: Tone control model.

5.2. Melody Control Model

5.2.1. Experimental Process. In order to evaluate the quality of the melody conversion results, three Beijing Opera pieces were selected for testing, followed by conversions using Only_dura, dura_F0, dura_SP, and all models, and Beijing operas produced by the four synthesis methods were compared with the original Beijing Opera. Among them, Only_dura uses only the duration control model for synthesis; dura_F0 uses only the base frequency control model and the duration control model for synthesis; dura_SP uses only the duration control model and the spectrum control model for synthesis; all models use three control models simultaneously. 'Real' is the source Beijing Opera.

So the melody control model can be summarized in Figure 16.

5.2.2. Experimental Results. The purpose of speech conversion is to make the converted speech sounds like the speech of a specific target person. Therefore, evaluating the performance of the speech conversion system is also based

Table 3: MOS grading.

MOS grading	
Score	MOS Evaluation
1	Uncomfortable and unbearable
2	There is a sense of discomfort, but it can endure
3	Can detect distortion and feel uncomfortable
4	Slightly perceived distortion but no discomfort
5	Good sound quality, no distortion

Table 4: Experimental results.

Experimental results			
ways	MOS fraction		
	BeiJing Opera1	BeiJing Opera2	BeiJing Opera3
Only_dura	1.25	1.29	1.02
dura_F0	1.85	1.97	1.74
dura_SP	1.78	2.90	2.44
all models	3.27	3.69	3.28
real	5	5	5

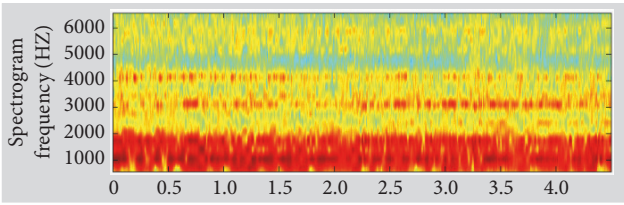

Figure 13: Source speech spectrogram.

on human-oriented auditory evaluation. In the existing subjective evaluation system, the MOS score test is an effective method for evaluating the voice quality, and the similarity test is a test method for judging the conversion effect of the system.

The MOS scoring criterion divides the speech quality into 5 levels; see Table 3. The tester listens to the converted speech and gives the score of the quality level to which the measured speech belongs according to these 5 levels. The MOS score is called the communication quality at about 3.5 minutes. At this time, the voice quality of the auditory reconstructed voice is reduced, but it does not prevent people from talking normally. If the MOS score is lower than 3.0, it is called synthetic speech quality. At this time, the speech has high intelligibility, but the naturalness is poor.

Find 10 testers and score MOS for the above composite results. The results are shown in Table 4.

5.3. Synthesis of Beijing Opera. Beijing Opera is mainly composed of words and melodies. The melody is determined by the pitch, tone, sound intensity, sound length, and other decisions. In this experiment, each word is distinguished by the unique characteristics of words such as zero-crossing rate and energy. Then the tone control model and the melody control model are designed and do extraction for important parameters of the fundamental frequency, spectrum, time, and so on, using MFCC, DTW, GMM, and other tools to analyze the extracted characteristic conversion and finally to the opera synthetic fragment.

Compared with other algorithms, the straight algorithm has better performance in terms of the natural degree of synthesis and the range of parameter modification, so the straight algorithm is also selected for the synthesis of the Beijing Opera.

Again, let the above-mentioned 10 testers perform MOS scoring on the above composite effect. The result is shown in Table 5.

According to the test results, it can be seen that the subjective test results reached an average of 3.7 points, indicating that the design basically completed the Beijing Opera synthesis. Although the Beijing Opera obtained by the synthesis system tends to originate in Beijing Opera, it is still acoustically different from the real Beijing Opera.

6. Conclusion

In this work, we have presented three novel generative models for Beijing Opera synthesis under the frame work of the straight algorithm, GMM and GAN. The objective metrics and the subjective user study show that the proposed models can achieve the synthesis of Beijing Opera. Given the recent enthusiasm in machine learning inspired art, we hope to continue our work by introducing more complex models and data representations that effectively capture the underlying melodic structure. Furthermore, we feel that more work could be done in developing a better evaluation metric of the quality of a piece; only then will we be able to train models that are

Table 5: Rating results.

MOS Score					
score	student1	student2	students student3	student4	student5
Source Opera fragment	5	5	5	5	5
Synthetic Opera fragment	4	4	4	3	3
score	student6	student7	students student8	student9	student10
Source Opera fragment	5	5	5	5	5
Synthetic Opera fragment	4	3	4	4	4

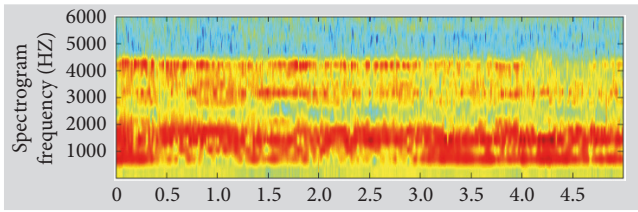

Figure 14: Target speech spectrogram.

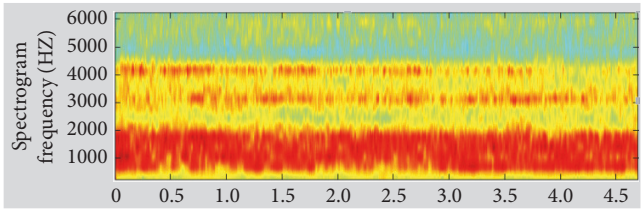

Figure 15: Converted speech spectrogram.

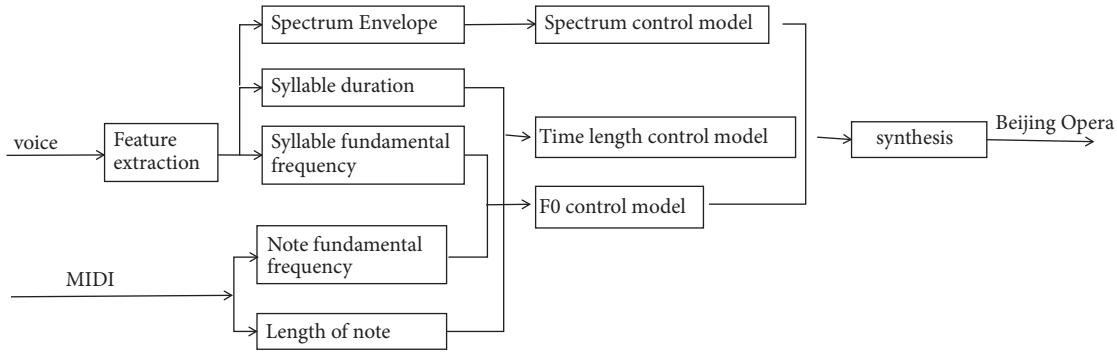

Figure 16: Melody control model.

truly able to compose the Beijing Opera singing art works with higher quality.

Data Availability

The [.wav] data of Beijing Opera used to support the findings of this study have been deposited in the [zenodo] repository: [http://doi.org/10.5281/zenodo.344932]. The previously reported straight algorithm used is available at http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.htmll. The code is available upon request from kawahara@sys.wakayama-u.ac.jp.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is sponsored by (1) the NSFC Key Funding no. 61631016; (2) the Cross Project "Research on 3D Audio Space and Panoramic Interaction Based on VR", no. 3132017XNG1750; and (3) the School Project Funding no. 2018XNG1857.

References

[1] D. Schwarz, "Corpus-based concatenative synthesis," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 92–104, 2007.

[2] J. Cheng, Y. Huang, and C. Wu, "HMM-based Mandarin singing voice synthesis using Tailored Synthesis Units and Question Sets," *Computational Linguistics and Chinese Language Processing*, vol. 18, no. 4, pp. 63–80, 2013.

[3] L. Sheng, *Speaker Conversion Method Research*, South China University of Technology doctoral dissertation, 2014.

[4] Y. Yang, *Chinese Phonetic Transformation System [Master's, thesis]*, Beijing Jiaotong University, 2008.

[5] S. Hasim and etal., "Fast and accurate recurrent neural network acoustic models for speech recognition," https://arxiv.org/abs/1507.06947.

[6] B. Tang, *Research on Speech Conversion Technology Based on GMM Model*, vol. 9, 2017.

[7] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, 2007.

[8] M. W. Macon, L. Jensen-Link, J. Oliverio, M. A. Clements, and E. B. George, "Singing voice synthesis system based on sinusoidal modeling," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP. Part 1 (of 5)*, pp. 435–438, April 1997.

[9] H. Gu and Z. Lin, "Mandarin singing voice synthesis using ANN vibrato parameter models," in *Proceedings of the 2008 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 3288–3293, Kunming, China, July 2008.

[10] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[11] C. Lianhong, J. Hou, R. Liu et al., "Synthesis of HMM parametric singing based on pitch," in *Proceedings of the 5th Joint Conference on Harmonious Human-machine Environment*, Xi'an, 2009.

[12] W. Wanliang and L. Zhuorong, "Advances in generative adversarial network," *Journal of Communications*, vol. 39, 2018.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," in *Proceedings of the Advances in neural information processing systems*, pp. 2672–2680, 2014.

[14] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," https://arxiv.org/abs/1511.06434.

[15] Interpretable representation learning by information maximizing generative adversarial nets,.

[16] A. Nguyen et al., "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," https://arxiv.org/abs/1605.09304.

[17] I. Goodfellow et al., *Deep Learning*, MIT Press, Cambridge, Mass, USA, 2016.