Development of the Music Teaching System Based on Speech Recognition and Artificial Intelligence

Zhang Zhichao

Lu Xun Academy of Arts, Yan’an University, Yan’an 716000, Shanxi, China

Correspondence should be addressed to Zhang Zhichao; 2020210425@mail.chzu.edu.cn

Received 27 January 2022; Revised 13 February 2022; Accepted 25 February 2022; Published 18 March 2022

Academic Editor: Muhammad Arif

Copyright © 2022 Zhang Zhichao. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Intelligent music teaching is the direction of subsequent reforms in music teaching methods. In order to improve the intelligence of music teaching, this paper conducts research on speech recognition technology. The speech conversion system based on Multiscale Star GAN extracts different levels of multiscale features of the global features of music utterance through the multiscale structure, which enhances the details of the converted speech and uses residual connections to alleviate the problem of gradient disappearance and enable the network to spread more deeply. In addition, after improving the speech recognition algorithm, this paper combines the needs of music teaching to construct a music teaching system based on speech recognition and artificial intelligence and designs system functional modules. Finally, this paper evaluates the performance of the system constructed in this paper by means of teaching experiments. From the experimental analysis, it can be seen that the system constructed in this paper has a good teaching effect.

1. Introduction

All-round development is the ideal of education that human beings strive to pursue, and music education contains a splendid style and long-standing characteristics that can help the educated to get closer to all-round development [1]. Therefore, we must make efforts to build a close connection between music and life in the process of learning music. Moreover, it is conducive to exploring and perfecting the auxiliary functions of music education and expanding the quality of life of young people in the music teaching classroom.

The impact of numerous music cultures, local government policy support, the employment of different versions of music appreciation textbooks, and the school’s right knowledge of music education all give directions and routes for the execution of auxiliary functions of music education. Simultaneously, the employment of “Internet +” and multimedia technologies in the classroom, under the complete stimulation of the information age, gives practical feasibility for the execution of today’s music discipline’s core literacy. As a result, the present scenario makes it possible to implement the auxiliary role of music education [2].

At the same time, the music education function is founded on the application of the music discipline’s core literacy against a foundation of core literacy. The auxiliary purpose of music education is to carry out the inheritance of music culture while also serving as an intermediate via its aesthetic function. It is possible for links to function. Music education can only develop the functions of moral education, intellectual education, and sports and can properly execute the essential features of music discipline, if it is founded on the aesthetic function of music. In this sense, educational concept renewal can promote the development and deepening of our music education’s auxiliary functions, contribute to the cultivation of people with a holistic development of morality, intelligence, and physical beauty, and enable the implementation of auxiliary functions of music education in teaching.

The long-term goal of the auxiliary function of music education is to mold the current students into people with core music literacy. Each lesson of music teaching needs to set a staged goal, which is based on the content of the
textbook and the physical and mental characteristics of the students. Staged goals can serve long-term goals and must be operable. Therefore, in addition to paying attention to the requirements of long-term goals, we must also pay attention to the unity, continuity, and gradual nature of the goals at each stage. Music education has the ability to stimulate wisdom, cultivate emotions, shape individuality, and realize human spiritual communication and integration. As an important part of general quality education, it is necessary to earnestly implement the spirit of relevant central documents, to comprehensively cultivate and develop ideal, ethical, educated, and disciplined socialist practical talents; moreover, it is necessary to fully develop the assistance of music education function to promote the cultivation of the comprehensive quality of ordinary students [3].

This paper uses speech recognition technology and artificial intelligence technology to study the development of music teaching system to further improve the reliability and stability of music teaching and to provide a theoretical basis for subsequent music teaching.

2. Related Work

There are different perspectives on the classification of speech conversion technology. Among them, a classification angle is to divide the speech conversion method into parallel text method and nonparallel text method according to whether the conversion model requires the same text or phoneme set. Parallel text refers to speeches with the same semantics but different speaker identities, while nonparallel text refers to speeches with different semantics and different speaker identities. Another classification angle is based on the language situation of the source and target speakers, which can be divided into same-language conversion and cross-language conversion [4]. Furthermore, depending on the number of speakers of the source and target speeches throughout the training process, speech conversion techniques may be split into one-to-one, many-to-one, and many-to-many speech conversion. Alternatively, it is separated into speech conversion in the training data set and speech conversion outside the training data set, depending on whether the speaker appears in the training data set. The majority of conventional speech conversion techniques are classified as speech conversion with parallel text, which means that the parallel text is utilised as training data to develop the mapping function or conversion model. Collecting parallel speeches, on the other hand, is time-consuming and labor-intensive in real-world circumstances, and it is not accessible at all in the fields of cross-language conversion and medical help. Furthermore, even if such parallel data is gathered, most speech conversion algorithms still need the training data to be time-aligned. To overcome the issue of temporal alignment problems, the alignment method always adds mistakes, necessitating more sophisticated processes such as exact corpus preparation or human correction [5].

Because of the problems with parallel text-to-speech conversion technology, nonparallel speech conversion technology has increasingly become the current standard. Some researchers have recommended that the model be assisted by background speech data [6] or a speech recognition module [7] to enable nonparallel speech conversion. Although background data and auxiliary modules aid in the resolution of nonparallel issues, they add to the complexity of the system and depend on auxiliary data. With the module’s overall performance, RBMs [8], variational autoencoders [9], and GANs are examples of probabilistic depth generation models that have been brought into the area of speech conversion in recent years [10]. These technologies do not rely on parallel data and the system performance is comparable to that of parallel speech conversion methods. The performance is equivalent. Among them, the voice conversion based on the cyclic uniform generation confrontation network [11] combines Cycle GAN [12] and gated convolutional neural networks (Gated CNNs) [13] to achieve voice conversion and obtain better experimental results. Although Cycle GAN-VC solves the dependence of the voice conversion method on parallel data and the voice quality and similarity after conversion are better, it still cannot achieve many-to-many voice conversion. Traditional voice conversion methods are mostly designed to learn one-to-one mapping, but it is very challenging to realize many-to-many voice conversion under nonparallel text. In order to solve the problem that Cycle GAN-VC cannot realize many-to-many speech conversion, the speech conversion based on star generation confrontation network (STARGAN-VC) [14] method introduces the speaker tag on the basis of Cycle GAN-VC. Realize nonparallel many-to-many speech conversion. Experiments show that the performance of this method is better than other current methods, such as many speech conversion methods based on VAE or other GANs [15].

Voice conversion methods based on VAE and GAN may directly avoid text alignment issues and give a new foundation for voice conversion in nonparallel text situations. Not only is the input feature encoded into a latent space feature in the VAE-based speech conversion series model, but the speaker is also encoded into a label feature. The aforementioned two characteristics are joined in the generating step to recreate the spectral feature. The most prevalent are the auxiliary classification variational autoencoder voice conversion (Auxiliary Classifier VAE VC, ACVAE-VC) [16], cyclic variational autoencoder voice conversion (Cycle VAE VC, Cycle VAE-VC) [17], and so on. In brief, the VAE-based model meets the criteria for parallel data and can do nonparallel conversion jobs; nevertheless, the output of the model’s decoder is too smooth, resulting in conversion in the speech conversion task. The post-voice quality is poor, and there is a buzzing sound, indicating that there are issues with the converted spectrum’s naturalness and resemblance. As a result, the voice conversion system based on the VAE framework does not function well. Through the learning of generators and discriminators, the GAN-based speech conversion system may improve the quality of synthesised speech. Literature [18] increased voice quality by proposing Cycle GAN-VC2, which incorporated a two-step counter loss to further improve spectrum generation quality. Literature [19] improved and improved the
original discriminator, with the capacity to extract features; literature [20] discovered that, in nonparallel data, the cross relationship between two speakers is utilized to perform speech conversion, but the resulting speech quality still has some pronunciation ambiguity and noise difficulties, the same as the above two. The same Cycle GAN-based approaches can only convert one-to-one, and there are still significant restrictions in practical applications. As a result, research into many-to-many nonparallel speech conversion has higher theoretical and practical utility.

3. Many-to-Many Speech Conversion Based on Star GAN under Nonparallel Text Conditions

Because Cycle GAN does not separate speaker identity features, it can only achieve one-to-one nonparallel speech conversion. If we want to achieve many-to-many, we can only train models with multiple speaker pairs. The Star GAN model was originally used on images to achieve many-to-many image conversion. Star GAN-VC separates speaker identity features on the basis of Cycle GAN and introduces many-to-one nonparallel speech conversion. If we want to achieve many-to-many, we can only train models with multiple speaker pairs. IQ—he Star GAN is composed of a generator $G$, a discriminator $D$, and a classifier $C$. The generator $G$ is composed of an encoding network and a decoding network. Unlike the Cycle GAN, the Star GAN uses only one generator $G$. The network structure diagram of Star GAN is shown in Figure 1.

As can be seen from Figure 1, Star GAN adds separately coded target label features to guide the generated target speaker’s speech. The generator $G$ encodes the input speech features of the source speaker to obtain the features on the semantic level and then inputs the tag codes of different speakers to decode and generate different speaker speech features. The generated features of the target speaker and the label of the source speaker are input into the generator $G$ again to obtain the reconstructed feature of the source speaker, and the loop loss $L_{\text{cyc}}(G)$ is calculated. In addition, the source speaker’s characteristics and the source speaker’s label are input into the generator $G$ to directly generate the source speaker’s speech features, and the feature mapping loss $L_{\text{id}}(G)$ is calculated [22].

$$L_{\text{cyc}}(G) = \mathbb{E}_{x \sim p(x|c_s), c_t \sim p(c_t), c_s \sim p(c_s)} \left[\|G(G(x_s, c_s), c_t) - x_s\|_1\right].$$ (1)

Among them, $x_s \sim p(x_s|c_s)$ represents the frequency spectrum characteristics of the source speaker, $c_t \sim p(c_t)$ represents the label of the source speaker, $c_s \sim p(c_s)$ represents the label of the target speaker, $G(G(x_s, c_s), c_t)$ represents the spectrum characteristics of the reconstructed source speaker, and $E_{x \sim p(x|c_s), c_t \sim p(c_t), c_s \sim p(c_s)}[\cdot]$ is the expected loss of the reconstructed source speaker’s spectrum and the true source speaker’s spectrum.

$$L_{\text{id}}(G) = \mathbb{E}_{c_t \sim p(c_t), x \sim p(x|c_s)} \left[\|G(x_s, c_t) - x_s\|_1\right].$$ (2)

Among them, $G(x_s, c_t)$ is the source speaker’s spectrum characteristics obtained after the source speaker’s frequency spectrum and source speaker label are input to the generator, and $E_{c_t \sim p(c_t), x \sim p(x|c_s)}[\cdot]$ is the expected loss of $x_s$ and $G(x_s, c_t)$.

The discriminator $D$ is used to judge whether the generated features conform to the feature distribution of the real speaker’s speech. The real speech and the label input of the speech speaker are input, and the probability that the speech feature conforms to the speech feature distribution of the real speaker is output. The pretraining generator distinguishes real speech features and generated speech features and minimizes the generator’s confrontation loss function $L_{\text{adv}}^D(G)$. The posttraining generator enables the generated speech features to deceive the discriminator into real speech and minimizes the generator’s confrontation loss function $L_{\text{adv}}^G(G)$.

The discriminator’s confrontation function is

$$L_{\text{adv}}^D(G) = -\mathbb{E}_{x \sim p(x|c_s), c_t \sim p(c_t)} \left[\log D(x_s, c_t)\right] - \mathbb{E}_{x \sim p(x|c_s), c_t \sim p(c_t)} \left[\log \left(1 - D(G(x_s, c_t), c_t)\right)\right].$$ (3)

Among them, $x_s \sim p(x_s|c_s)$ represents the spectrum characteristics of the target speaker, $D(x_s, c_t)$ represents the discriminator $D$ to determine whether the real spectrum characteristics are true, $G(x_s, c_t)$ represents the spectrum characteristics of the target speaker generated by the
generator $G$, and $D(G(x_s, c_t), c_i)$ represents whether the spectrum characteristics generated by the discriminator are true, $E_{x \sim p(x|c_i)} c_i \sim p(c_i)$ represents the expectation of conforming to the probability distribution of the true feature, and $E_{x \sim p(x|c_i)} c_i \sim p(c_i)$ represents the expectation of conforming to the probability distribution of the generated feature.

The confrontation loss function of generator is

$$L_{adv}^G (G) = -E_{x \sim p(x|c_i)} c_i \sim p(c_i) [\log D(G(x_s, c_t), c_i)]. \quad (4)$$

Among them, $G(x_s, c_t)$ represents the spectrum characteristics generated by the generator, $D(G(x_s, c_t), c_i)$ represents whether the generated spectrum characteristics are true by the discriminator, and $E_{x \sim p(x|c_i)} c_i \sim p(c_i)$ represents the expectation of the probability distribution of the generated feature.

The classifier is used to determine which speaker the generated speech belongs to. First, the classifier is trained with real speaker’s speech features, and the classification loss $L_{cls}^G (C)$ of classifier is calculated. Furthermore, the classifier is used to judge the generated speaker’s speech category, calculate the classification loss $L_{cls}^G (G)$ of the generator, and train the generator.

The classification loss of the classifier is

$$L_{cls}^G (C) = -E_{x \sim p(x|c_i)} c_i \sim p(c_i) [\log P_C (c_i|x_s)]. \quad (5)$$

The classification loss of the generator is

$$L_{cls}^G (G) = -E_{x \sim p(x|c_i)} c_i \sim p(c_i) [\log P_C (c_i|G(x_s, c_t))]. \quad (6)$$

Among them, $P_C (c_i|x_s)$ represents the probability that the classifier distinguishes the target speaker’s spectrum characteristics label as the target speaker’s label $c_i$. $P_C (c_i|G(x_s, c_t))$ represents the probability that the classifier discriminates that the target speaker’s spectrum characteristics label is the target speaker label $c_i$, and $G(x_s, c_t)$ represents the target speaker’s spectrum characteristics generated by the generator.

The generator’s cycle consistency loss and feature mapping loss, the generator’s classification loss, and the generator’s confrontation loss are integrated. The total loss function of the generator is

$$L_G (G) = L_{adv}^G (G) + \lambda_{cls} L_{cls}^G (C) + \lambda_{cyc} L_{cyc} (G) + \lambda_{id} L_{id} (G). \quad (7)$$

Among them, $\lambda_{cls} \geq 0, \lambda_{cyc} \geq 0, \lambda_{id} \geq 0$ is a regularization parameter, which, respectively, represents the weights of classification loss, cycle consistency loss, and feature mapping loss, and $L_{adv}^G (G)$, $L_{cls}^G (C)$, $L_{cyc} (G)$, $L_{id} (G)$ represent the generator’s confrontation loss, generator’s classification loss, cycle consistency loss, and feature mapping loss, respectively.

The Star GAN-based speech conversion system includes two stages of training and conversion, as shown in Figure 2.

In the training phase, through the WORLD speech analysis/synthesis model, the spectrum characteristics $x$, fundamental frequency features $\log f_0$, and aperiodic features of each speaker’s sentence are extracted. Then, the spectrum characteristics $x_t$ of the source speaker, the spectrum characteristics $x_s$ of the target speaker, the source speaker label $c_t$, and the target speaker label $c_s$ are input into the Star GAN for training. In the training process, it is necessary to make the loss function of the generator, the loss function of the discriminator, and the loss function of the classifier as small as possible until the set number of iterations, so as to obtain a trained Star GAN. The fundamental frequency conversion function from the fundamental frequency of the source speaker’s speech to the fundamental frequency $f_{0t}$ of the target speaker’s speech is constructed.

In the conversion stage, the source speaker’s speech in the corpus to be converted is used to obtain the spectral envelope characteristics $x'_t$, aperiodic features, and fundamental frequency $\log f_{0t}$ through the WORLD speech analysis/synthesis model. Then, the source speaker’s spectral envelope characteristics $x'_t$ and target speaker’s label characteristics $c'_t$ are input into the trained Star GAN to reconstruct the target speaker’s spectral envelope characteristics $x'_{tc}$. Then the fundamental frequency of the source speaker’s speech $f_{0t}$ is converted into the fundamental frequency $\log f'_{0t}$ of the target speaker through the trained

---

**Figure 2:** The training and conversion flowchart of Star GAN.
fundamental frequency conversion function. Finally, the target speaker’s spectral envelope characteristics $x_{t\sigma}$, the fundamental frequency, and aperiodic characteristics of the target speaker are synthesised through the WORLD speech analysis/synthesis model to obtain the converted speaker’s speech.

The model consists of three parts: a generator $G$ that generates the spectrum of the target speaker, a discriminator $D$ that determines whether the input is a real spectrum or a generated spectrum, and a classifier $C$ that determines whether the label of the spectrum belongs to the speaker. Among them, the generator $G$ is composed of a precoding network, a multiscale module, and a decoding network. The network structure diagram of multiscale Star GAN is shown in Figure 3.

The speech feature $x_s$ and target speaker label $c_t$ are input into the precoding network of the generator $G$ containing the multiscale module to obtain the global characteristics of the target speaker domain. After that, the global features are imported into the multiscale module. Through the multiscale module, the range of the network’s receptive field is increased, and features of different levels are more fully extracted. In the multiscale module, the global feature $G(x_s)$ is divided into $s$ feature map subsets. Each feature map subset represents a scale and level of features, from the perception information of feature map subsets of different scales. Moreover, residual connections are used in the input and output of multiscale to build a hierarchical connection realization and alleviate the problem of gradient disappearance. The $s$ feature map subsets are spliced to obtain hierarchically connected multiscale features $G_M(x_s)$, which are input to the decoder of the generator to obtain the spectrum characteristics $x_{s\sigma}$ of the target speaker.

The spectrum characteristics $x_{s\sigma}$ of the target speaker and the label characteristics $c_s$ of the source speaker are reinput to the precoding network of the generator $G$ containing the multiscale module to obtain the global characteristics $G(x_s)$ of the source speaker domain. After that, the hierarchical connected multiscale feature $G_M(x_s)$ is obtained through multiscale coding. It is input to the decoder of the generator to obtain the spectrum characteristics $x_{sc}$ of the reconstructed source speaker, thereby obtaining the cycle loss of the spectrum characteristics $x_s$ of the source speaker and the spectrum characteristics $x_{sc}$ of the reconstructed source speaker.

The cyclic consistent loss function is

$$L_{cyc}(G) = E_{c_s \sim p(c_s), c_t \sim p(c_t)} [\|G(G(x_s, c_s), c_t) - x_s\|_1].$$

(8)

Among them, $x_s \sim p(x_s | c_s)$ represents the spectrum characteristics of the source speaker, $c_t \sim p(c_t)$ represents the label of the source speaker, and $c_s \sim p(c_s)$ represents the label of the target speaker, $G(G(x_s, c_s), c_t)$ represents the frequency spectrum characteristics of the reconstructed source speaker, and $E_{x_s \sim p(x_s | c_s), c_s \sim p(c_s), c_t \sim p(c_t)} [\cdot]$ is the expected loss of the reconstructed source speaker’s spectrum and the real source speaker’s spectrum.

In addition, the source speaker’s spectrum characteristics $x_s$ and source speaker’s label characteristics $c_s$ are input to the generator to obtain the source speaker’s spectrum characteristics $x_{s\sigma}$ after feature mapping, so as to obtain the generator feature mapping loss $L_{id}$. The feature mapping loss function is

$$L_{id}(G) = E_{c_s \sim p(c_s), x_s \sim p(x_s | c_s)} [\|G(x_s, c_s) - x_s\|_1].$$

(9)

Among them, $G(x_s, c_s)$ is the source speaker spectrum characteristics obtained after the source speaker’s spectrum and source speaker’s label are input to the generator, and $E_{c_s \sim p(c_s), x_s \sim p(x_s | c_s)} [\cdot]$ is the expected loss of $x_s$ and $G(x_s, c_s)$. The discriminator is used to determine if the actual/generated feature conforms to the feature distribution of the real speaker’s speech, and it outputs the probability that the speech feature conforms to the real speaker’s speech feature distribution. The enhanced discriminator just accepts real/generated speech and no longer requires the speech speaker’s tag; thus the source speaker’s actual speech is used as the real speech input, and the produced target speaker’s speech is used as the false speech input discriminator. The pretraining discriminator distinguishes real speech features and generated speech features and minimizes the discriminator’s confrontation loss function $L_{adv}^G(G)$. The posttraining discriminator enables the generated speech features to deceive the discriminator into real speech and minimizes the discriminator’s confrontation loss function $L_{adv}^G(G)$. 

![Network structure diagram of multiscale Star GAN.](image)
The confrontation loss function of the discriminator is

\[ L_{adv}^D (D) = -E_{x \sim p(x), c \sim p(c)} [\log D(x)] - E_{x \sim p(x), c \sim p(c)} [\log (1 - D(G(x, c)))], \]  

(10)

4. Music Teaching System Based on Speech Recognition and Artificial Intelligence

On the basis of the above analysis, a music teaching system based on speech recognition and artificial intelligence is constructed. The system frame structure is shown in Figure 4.

The basic music theory knowledge display submodule, the music basic theory knowledge retrieval submodule, and the music basic theory knowledge query submodule are all part of the basic music theory knowledge learning module. The display submodule is the system’s principal module for supporting basic music theory learning, and it may show knowledge items and specialised material related to basic music theory. The retrieval and query submodule allows you to search for and retrieve relevant information. Figure 5 depicts the music teaching system’s system function flowchart, which is based on voice recognition and artificial intelligence.

The detailed design of the user management module is shown in Figure 6:

Within the user administration module, the logical link is rather straightforward. After passing the login verification, the system administrator may go straight to the system administration page and utilise the features of adding, changing, and removing users. If the login verification fails, the system switches to a reverification interface. The system administrator category, user management form category, system administrator information category, teacher user information category, student user information category, and process control category are the key categories defined by the module.

5. Performance Detection of the Music Teaching System Based on Speech Recognition and Artificial Intelligence

After constructing a music teaching system based on speech recognition and artificial intelligence technology, the performance of the system constructed in this paper is tested. According to the actual needs and requirements of the construction of the system in this paper, the performance of
Figure 6: Function flowchart of the user management module.

Table 1: Statistical table of speech recognition effect of the music teaching system.

| Number | Speech recognition | Feature extraction | Number | Speech recognition | Feature extraction |
|--------|--------------------|--------------------|--------|--------------------|--------------------|
| 1      | 92.4               | 66.5               | 31     | 90.0               | 67.8               |
| 2      | 83.7               | 67.0               | 32     | 83.7               | 67.3               |
| 3      | 92.6               | 66.4               | 33     | 90.2               | 66.6               |
| 4      | 87.6               | 66.4               | 34     | 89.6               | 62.9               |
| 5      | 88.7               | 69.9               | 35     | 87.0               | 67.0               |
| 6      | 88.3               | 65.2               | 36     | 85.3               | 62.8               |
| 7      | 88.1               | 61.3               | 37     | 91.5               | 67.6               |
| 8      | 83.4               | 62.6               | 38     | 83.5               | 69.4               |
| 9      | 92.3               | 69.2               | 39     | 81.8               | 62.7               |
| 10     | 84.5               | 63.0               | 40     | 86.7               | 67.5               |
| 11     | 86.0               | 69.7               | 41     | 82.2               | 64.3               |
| 12     | 92.5               | 63.9               | 42     | 86.5               | 64.2               |
| 13     | 80.7               | 64.9               | 43     | 86.3               | 62.4               |
| 14     | 80.2               | 62.9               | 44     | 83.5               | 67.9               |
| 15     | 82.0               | 65.9               | 45     | 82.1               | 68.4               |
| 16     | 82.3               | 63.6               | 46     | 90.2               | 68.7               |
| 17     | 81.5               | 63.3               | 47     | 91.0               | 65.8               |
| 18     | 83.7               | 62.8               | 48     | 84.1               | 69.7               |
| 19     | 81.9               | 63.1               | 49     | 82.3               | 63.5               |
| 20     | 80.7               | 62.8               | 50     | 84.6               | 68.4               |
| 21     | 83.5               | 69.1               | 51     | 81.0               | 64.7               |
| 22     | 84.9               | 63.0               | 52     | 90.9               | 64.8               |
| 23     | 89.2               | 61.7               | 53     | 88.7               | 63.6               |
| 24     | 82.0               | 66.5               | 54     | 92.5               | 61.3               |
the system is tested from two aspects: speech recognition and system teaching effect.

This paper uses 60 students from a university as a research sample to design experiments to analyze the performance of music teaching system. First, this paper analyzes the speech recognition effect of the system constructed in this paper, and the results are shown in Table 1 and Figure 7.

| Number | Speech recognition | Feature extraction | Number | Speech recognition | Feature extraction |
|--------|--------------------|--------------------|--------|--------------------|--------------------|
| 25     | 80.3               | 64.7               | 55     | 86.1               | 69.5               |
| 26     | 85.0               | 67.3               | 56     | 87.5               | 67.1               |
| 27     | 91.6               | 61.0               | 57     | 86.7               | 65.4               |
| 28     | 91.6               | 64.6               | 58     | 80.8               | 61.4               |
| 29     | 82.5               | 61.1               | 59     | 79.1               | 61.7               |
| 30     | 91.3               | 68.5               | 60     | 90.7               | 67.4               |

Table 2: Statistics table of music teaching effect evaluation of the music teaching system.

| Number | Teaching effect | Number | Teaching effect | Number | Teaching effect |
|--------|-----------------|--------|-----------------|--------|-----------------|
| 1      | 72.3            | 21     | 78.8            | 41     | 88.1            |
| 2      | 82.2            | 22     | 87.6            | 42     | 84.0            |
| 3      | 85.7            | 23     | 74.0            | 43     | 77.3            |
| 4      | 78.4            | 24     | 71.4            | 44     | 71.9            |
| 5      | 74.5            | 25     | 70.5            | 45     | 74.1            |
| 6      | 65.2            | 26     | 87.1            | 46     | 80.2            |
| 7      | 72.1            | 27     | 68.2            | 47     | 80.5            |
| 8      | 84.1            | 28     | 89.1            | 48     | 80.5            |
| 9      | 71.8            | 29     | 77.4            | 49     | 72.1            |
| 10     | 70.7            | 30     | 75.3            | 50     | 72.4            |
| 11     | 82.8            | 31     | 82.1            | 51     | 73.7            |
| 12     | 78.7            | 32     | 71.1            | 52     | 65.4            |
| 13     | 81.3            | 33     | 67.9            | 53     | 77.9            |
| 14     | 87.7            | 34     | 86.3            | 54     | 85.8            |
| 15     | 89.3            | 35     | 87.6            | 55     | 85.6            |
| 16     | 74.2            | 36     | 88.6            | 56     | 82.8            |
| 17     | 85.0            | 37     | 68.4            | 57     | 71.2            |
| 18     | 75.2            | 38     | 86.7            | 58     | 79.8            |
| 19     | 68.3            | 39     | 80.0            | 59     | 66.1            |
| 20     | 65.7            | 40     | 78.1            | 60     | 66.5            |
Teaching effect

![Figure 8: Statistics diagram of music teaching effect evaluation of the music teaching system.](image)

The above results show that the speech recognition ability of the music teaching system constructed in this paper is better. On this basis, this paper uses the teaching scoring method to verify the teaching performance of the music teaching system constructed in this paper. The results obtained are shown in Table 2 and Figure 8.

Through the above experimental results, it can be seen that the music teaching system based on speech recognition and artificial intelligence constructed in this paper has greater advantages than the traditional teaching mode.

6. Conclusion

The maturation of computer network technology, as well as the steady use of campus networks, has laid a solid platform for the informatization of music education and instruction. Computer and network technologies may help with music education and teaching in ways that conventional classroom education and teaching cannot. As a result, an online music education and teaching assistant system that matches the requirements of modern educational technology development is required. In terms of the relevance of instructional system design theory, the essential position of instructional system design theory in the instructional design discipline is explained via in-depth discussion. Furthermore, the parallels and contrasts between teaching system design theory and teaching theory are examined in this work. Furthermore, this research designs a music education system based on voice recognition and artificial intelligence to validate the system’s performance using speech recognition technology. The results of the tests suggest that the approach proposed in this research is successful in music instruction.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] S. Orlandi, C. A. R. Garcia, A. Bandini, and G. Donzelli, C. Manfredi, “Application of pattern recognition techniques to the classification of full-term and preterm infant cry,” *Journal of Voice*, vol. 30, no. 6, pp. 656–663, 2015.
[2] C.-C. Hsu, K.-M. Cheong, T.-S. Chi, and Y. Tsao, “Robust voice activity detection algorithm based on feature of frequency modulation of harmonics and its DSP implementation,” *IEICE-Transactions on Info and Systems*, vol. E98.D, no. 10, pp. 1808–1817, 2015.
[3] P. H. Kumar and M. N. Mohanty, “Efficient feature extraction for fear state analysis from human voice,” *Indian Journal of Science & Technology*, vol. 9, no. 38, pp. 1–11, 2016.
[4] R. Rhodes, “Aging effects on voice features used in forensic speaker comparison,” *International Journal of Speech Language and the Law*, vol. 24, no. 2, pp. 177–199, 2017.
[5] C. T. Herbst, S. Hertegard, D. Zanger-Borch, and P. A. Lindestad, “Freddie mercury—acoustic analysis of speaking fundamental frequency, vibrato, and sub-harmonics,” *Logopedics Phoniatrics Vocology*, vol. 42, no. 1, pp. 1–10, 2016.
[6] J. Al-Tamimi, “Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: implications for formal representations,” *Laboratory Phonology*, vol. 8, no. 1, pp. 1–40, 2017.
[7] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.
[8] C. Kim and R. M. Stern, “Power-normalized cc) for robust speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, 2016.
[9] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, “Audio-visual speech recognition using deep learning,” *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.
[10] Y. Qian, M. Bi, T. Tan, and K. Yu, ”Very deep convolutional neural networks for noise robust speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.
[11] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, ”An overview of noise-robust automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.
[12] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, ”Automatic speech recognition for under-resourced languages: a survey,” *Speech Communication*, vol. 56, no. 3, pp. 85–100, 2014.
[13] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, ”Hybrid CTC/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
[14] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, ”An analysis of environment, microphone and
data simulation mismatches in robust speech recognition,” *Computer Speech & Language*, vol. 46, no. 3, pp. 535–557, 2017.

[15] P. Swietojanski, A. Ghoshal, and S. Renals, “Convolutional neural networks for distant speech recognition,” *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[16] Q. K. Duong and H. T. Duong, “A review of audio features and statistical models exploited for voice pattern design,” *Computer Science*, vol. 3, no. 2, pp. 36–39, 2015.

[17] M. Sarria-Paja, M. Senoussaoui, and T. H. Falk, “The effects of whispered speech on state-of-the-art voice based biometrics systems,” in *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, no. 1, pp. 1254–1259, Nova Scotia, Canada, May 2015.

[18] A. Leeman, H. Mixdorff, M. O’Reilly, M.-J. Kolly, and V. Dellwo, “Speaker-individuality in Fujisaki model f0 features: implications for forensic voice comparison,” *International Journal of Speech Language and the Law*, vol. 21, no. 2, pp. 343–370, 2015.

[19] A. K. Hill, R. A. Cárdenas, J. R. Wheatley et al., “Are there vocal cues to human developmental stability? relationships between facial fluctuating asymmetry and voice attractiveness,” *Evolution and Human Behavior*, vol. 38, no. 2, pp. 249–258, 2017.

[20] K. Chapi, V. P. Singh, A. Shirzadi et al., “A novel hybrid artificial intelligence approach for flood susceptibility assessment,” *Environmental Modelling & Software*, vol. 95, no. 1, pp. 229–245, 2017.

[21] M. R. Hashemi, M. L. Spaulding, A. Shaw, H. Farhadi, and M. Lewis, “An efficient artificial intelligence model for prediction of tropical storm surge,” *Natural Hazards*, vol. 82, no. 1, pp. 471–491, 2016.

[22] T. Sustrova, “A suitable artificial intelligence model for inventory level optimization,” *Trends Economics and Management*, vol. 10, no. 25, pp. 48–55, 2016.