




Optimization of Multimedia Computer-aided Interaction System of Vocal Music Teaching Based on Voice Recognition

Meili Zhou¹  and Tianzhuo Gong^{2,3*} 

¹Art College, Heilongjiang University, Harbin, 150000, China, belcanto0451@sohu.com

²Music Collage, Capital Normal University, Beijing, 100000, China, dragonever@sina.cn

³Music Collage, Harbin Normal University, Harbin, 150000, China, dragonever@sina.cn

Corresponding author: Tianzhuo Gong, dragonever@sina.cn

Abstract. Based on the computer-aided system structure and combining the characteristics of the music sight-singing subject in the teaching content and teaching methods, this paper studies the vocal music teaching system model of sound recognition and completes the design of the overall system architecture, combining music sight-singing. The characteristics of the subject in the music teaching system provided a theoretical basis and reference for the realization of the system. It formed a complete and feasible enhanced string instrument and computer interactive electronic music system framework to meet the teaching needs of experimental interactive electronic music creation and performance. A set of scoring difficult feature sets with good discrimination is proposed, and a score recommendation algorithm and strategy based on score difficulty is designed to realize the personalized push of learning resources in intelligent tutoring. We put forward a method and model for evaluating the level of music sight-singing ability with certain applicability. Degree defines multiple ability evaluation indicators to realize the intelligent evaluation of the learner's cognitive level in the intelligent guidance system. This study provides a solid foundation for further research in the future.

Keywords: Voice Recognition; Vocal Music Teaching; Multimedia Computer-Aided; Interactive System

DOI: <https://doi.org/10.14733/cadaps.2021.S2.113-122>

1 INTRODUCTION

Personalized teaching methods are the key to improving the teaching level and training innovative talents. Due to the limited teaching resources and teachers, it is difficult to implement one-to-one personalized teaching for each student in the traditional teaching environment. With the rise of Internet teaching, the limitations of traditional teaching methods in terms of time and space have been resolved [1]. The web-based teaching platform builds visualized teaching scenarios and teaching resources, making teaching behaviors no longer limited to classrooms and classrooms. Students can learn freely through the online learning platform without time limit, location, and

freedom. Liu [2] showed that the various services provided by the platform for learners to provide one-to-one guidance and help for learners like private teachers. Hinton [3] pointed that learners with learning resources and learning environment can truly achieve personalized teaching, but also requires the learning platform to actively interact with learners. We need to understand the learner's personalized features, and imitate the teaching methods of teachers to provide different learners with targeted the intelligent teaching system which is a learning support system designed for realizing this intelligent teaching method. Dahl [4] said that music sight-singing subject is based on mastering the basic principles of music and imagination of music. It is all engaged in music. One of the basic skills that a musician's musician masters has always been valued by those engaged in music professional education. With the continuous development of the music education system, the teaching ideas and teaching methods of solfeggio are constantly innovating. However, teaching methods have been using classroom teaching and training methods for many years. This boring teaching method is not only inefficient but also greatly influential. Students are interest in learning lacks for the initiative to learn. Many words are used in the teaching of sight-reading teaching. This software can only assist teaching. The existing online courses and online learning platforms only provide learners with rich learning. Resources and different learning paths cannot meet the need of intelligent teaching.

Mohamed [5] developed large glass organs used light to reflect spectral characteristics. In 2015, the MIT Multimedia Laboratory tried to visualize the achievements in the field of music information retrieval in the project and used signal processing to extract information such as pitch, chords, and emotions in music. Nowadays, people can also see glowing music devices such as flames, light, Tesla coils, etc. Netzer [6] carried out the research and design of the color music system based on the fast transform. Dehak [7] conducted an in-depth discussion on the multi-channel mapping mode to enhance the human nature in the visual work. Comparing the international situation, China's relevant have reached the international level, but it is undeniable that most of the equipment and software tools used are developed by European and American researchers. The scale and diversity of the research are also quite different from abroad. The level of art and science disciplines needs to be further improved. In recent years, it has also begun to pay attention to this form of music information visualization, and researchers have studied the color performance in animation viewing. China Academy of Art introduced the visual presentation of music content in the age of digital media. Bojanowaski [8] took abstract animation as the starting point, and studied the visual expression of music from the perspective of animation creation. We advocated the design in combination with modern interactive technology, which is forward-looking. This kind of research mostly comes from the professional fields of fine arts, animation, film, and television, etc. It is more focused on the analysis of the visual angle, but the mining of musical information is less, and the lack of more systematic acoustic principle support [9]. This research is based on the theoretical research results of the existing intelligent guidance system, combined with the disciplinary characteristics of music sight-singing, in-depth study of the model and design method in the music sight-singing intelligent guidance system. This study has accumulated a lot for the realization of the intelligent guidance system. The algorithm proposed in this paper provides a feasible solution for the problems of personalized teaching resource push and ability evaluation in music sight-singing. It has the positive significance of further perfecting music sight-singing personalized teaching.

2 DESIGN AND ANALYSIS OF VOICE RECOGNITION INTERACTIVE SYSTEM

2.1 Audio Signal Processing System Analysis

The similarity is an index that comprehensively evaluates the similarity between two things. The similarity is a selection criterion widely used in recommendation systems based on neighborhood collaborative filtering algorithms. Based on the neighborhood, it is recommended based on adjacent elements, and the selection process of adjacent elements is the calculation process of similarity. Elements in different scenes have different quantization standards for their similarity.

The distance between points in space is a measure of similarity. The closer the vectors in space are, the more similar they are. According to different measurement standards, the calculation method of similarity is mainly divided into two categories. One is a distance measurement, such as Euclidean distance, Manhattan distance, and the other is the related measurement.

The data obtained by the above has various sensors through the single-chip microcomputer which is unified into the Max/MSP interactive design platform for processing. It has integrating all available variables and trigger data, and various interactive electronic music function modules, artistically Interactive electronic music creation. We designed the parameter feedback user interface used in the performance. If some variable parameters are docked with the real-time video generation program, multimedia interactive electronic music creation with digital images can be completed. Figure 1 shows the overall design of the system signal flow.

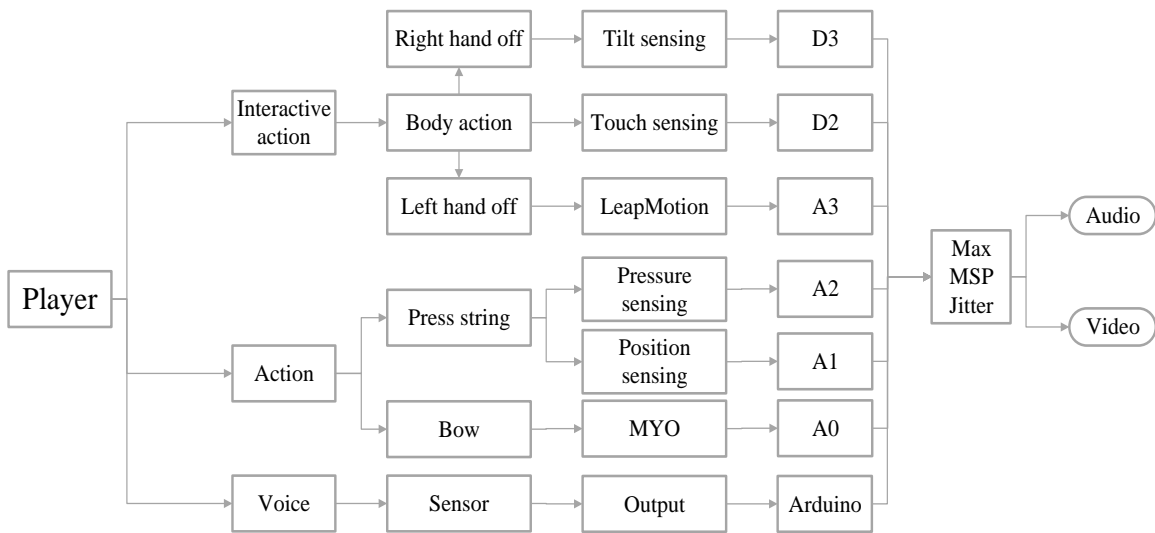


Figure 1: Signal flow chart of the music performance system.

The design of new electronic musical instruments needs to utilize sensing device detection and feedback DMIs, which can be included in the musical expression of the new interface. The new interface for musical instrument expression is a new term that is familiar, and there is not even an accurate Chinese translation recognized by the academic community. It refers to the physical properties of traditional acoustic or electro-acoustic instruments. It refers to the use of sensing devices and signal transmission methods to finally digitize the variables generated by various physical parameters of performance. It used the data stream to drive a variety of the algorithm performs interactive electronic music creation and performance. The sensor occupies an important position at the beginning of the data in this process. The sensor is reasonable determines whether the physical dynamics in the performance can be detected correctly and steadily. The beginning of the data stream is the basis for the subsequent interactive electronic music module to be correctly driven.

From the perspective of the new interface design related to enhanced strings instruments, combined with all the sensory categories that can be produced by the performance above and the common sensor types described in Table 1, the sensor design of this system can refer to the selection of the following types of sensors.

Describe the actions of a strung instrument	Detection sensor type
---	-----------------------

Left hand pressing string pressure	FSR pressure sensor
Right-hand bow finger pressure	FSR pressure sensor
Right-hand bow position	IR infrared or Ultrasound ultrasonic ranging sensor
Speed state of right-hand bow	Accelerator acceleration sensor
Right-hand bow angle change	Gyroscope
Body swing	Tilt sensors/mercury switches
Finger commands such as start/end/trigger	Tactile sensors
Separate spatial movement of the left and right hands	Leap Motion infrared 3D distance sensor
Other space actions	Video/Light video/light sensor
Playing sound	Microphone

Table 1: List of sensor category selection schemes for musical instrument performance.

The pressure sensor is essentially a resistance that changes with pressure. The resistance changed data cannot be directly detected, received, or collected by the single-chip microcomputer, but needs to be converted into a voltage change signal. One of the conversion methods is to load a 5V single-chip voltage with a single-chip experimental breadboard. It connected a pressure sensor and a 10K Ω pull-up resistor in series, and connect the analog signal line in parallel to the negative electrode to measure the voltage. The second conversion method is to use the officially equipped RFP-ZHII resistor-voltage conversion module can easily variable resistance pressure directly into the analog change of the output voltage, the voltage changes are consistent with changes in the resistance equation. The resistance-voltage conversion module has another important function. It outputs a low-voltage trigger signal immediately when it detected that the pressure reaches the maximum value. Therefore, a combination of a pressure sensor can be used to obtain a continuously changing pressure parameter and a maximum trigger signal.

2.2 System Evaluation Analyses

According to the relevant principles of pedagogy, there are three main relationships between the knowledge points of the discipline, such as interdependence, parent-child relationship, and mutual independence. Dependency was defined according to the process of knowledge acquisition. The parent-child relationship is the inclusion relationship between knowledge. There is a parent-child relationship between the knowledge of the child nodes. There is no dependency and inclusion between the knowledge points of the parallel relationship, the two are independent of each other in knowledge structure and knowledge acquisition, and do not affect each other. When constructing the knowledge model, the dependency relationship and the parent-child relationship need to be manifested explicitly, and the structure described the knowledge point should include two attributes of dependent knowledge and parental knowledge.

The model can be built according to the structure of the knowledge tree. The nodes in the subject knowledge point three are collectively called knowledge points. The leaf nodes are the smallest knowledge points that can be learned. The parent node usually contains multiple meta-knowledges, so it is called a compound knowledge point. The knowledge element further comprises a resource property knowledge of the resources used to bind the teaching point. It composited knowledge points which needed to include meta-knowledge point node attributes, recording all child nodes to facilitate knowledge retrieval. As shown in Figure 2, it is the structure of teaching content.

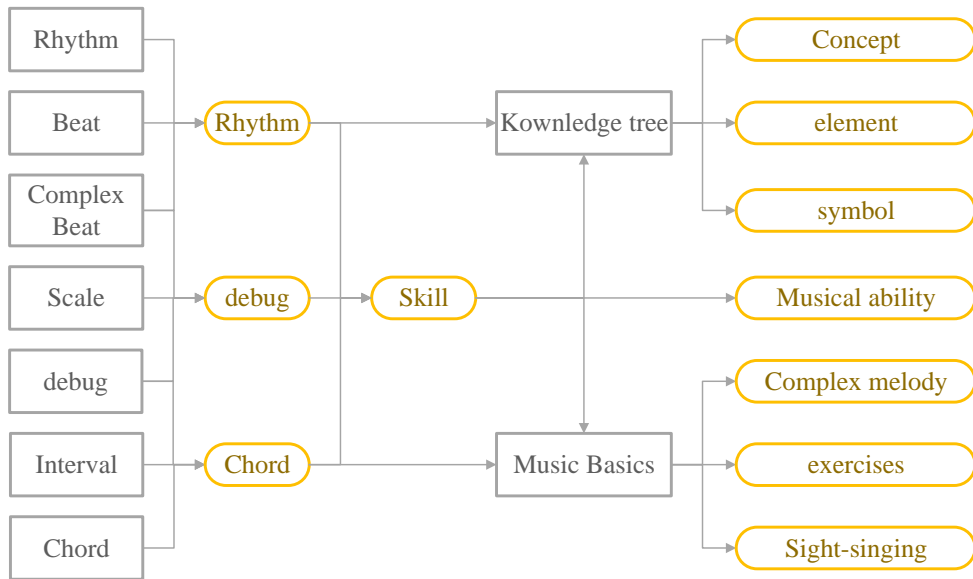


Figure 2: Teaching content structure.

The recommendation of learning resources based on the ability level is mainly to examine the difficulty of learning resources, that is, the degree of adaptation of the knowledge contained in the resources and the ability of the learner. Pushing knowledge for learners with lower abilities is difficult and complicated, and the learners will lose their interest in learning because of frustration; learning content pushed for learners with higher abilities is too simple, and it is difficult to achieve learner abilities Effective improvement. To ensure that the learning resources recommended by the teacher model, this paper proposes a learning resource recommendation algorithm based on the difficulty of the score. As decision support for resource recommendation in the teacher model, the algorithm-related content will be introduced in subsequent chapters. In the learning process of sight-singing, there are two parts, sight-singing practice and sight-singing test, which need to recommend resources for learners. To improve learning efficiency in sight-singing practice, it is necessary to recommend suitable practice resources for learners before starting the practice and after completing the scoring practice. Before learning, we analyze the sight-singing ability indicators which included in the learning content selected by the learner. the learner used the level value as the difficulty corresponding to the score to be selected. To the ability indicators for the rhythm pattern, the learner ability level is 0.2, and it selected from the score associated with the difficulty of rhythm pattern feature value of 0.2. It chooses a score for learners according to the difficulty recommendation algorithm and start learning. The learner continues to learn the same knowledge after practicing the current score or the practice time is too long in the current score. The recommendation system will also automatically recommend scores for them, querying the learner's average score in the current score practice. The score value is multiplied by the difficulty feature value of the current score, and the calculation result is used as the difficulty feature of the score to be recommended for music which is difficult feature vector. The average score is 0.6. The sight-singing exercises include score of 1 to 6 knowledge points of sight-singing skills, generally, the phrases or no more than 8 measures. The number of knowledge points is not limited. Such resources are not tied to knowledge points, but the knowledge points of sight-singing skills included, such as the number of rhythm patterns, intervals, tonality, and range, should be accurately recorded [10]. At the same time, to facilitate the recommendation of resources, it is necessary to extract the difficult feature of the score according to the method. The same files format and sample type resources, particularly as defined in Figure 3. A longer score or a complete musical composition contains multiple knowledge points. Compared with skill learning resources,

there is only a difference in duration and complexity of the score, so the model structure is the same.

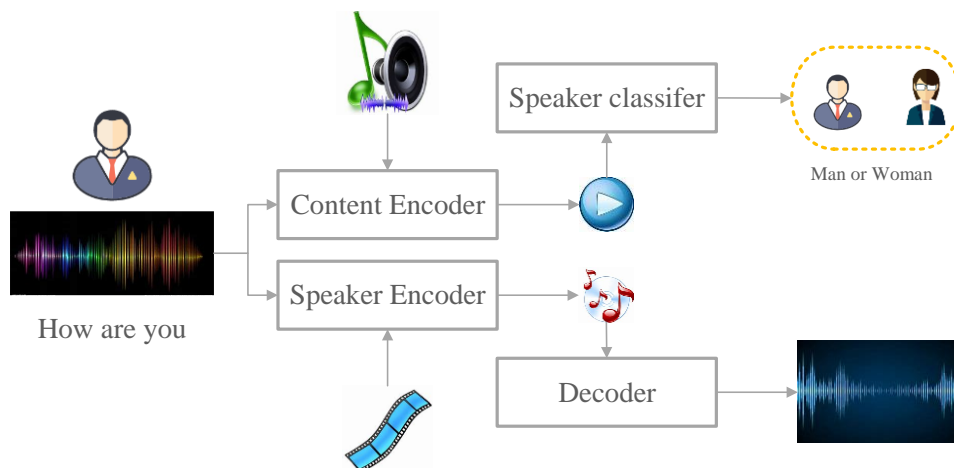


Figure 3: Flow chart of the score recommendation algorithm.

3 RESULTS AND ANALYSIS

3.1 Voice Recognition Difficulty Analyses

To verify the effectiveness of the features defined in this article in distinguishing the difficulty of the score, it must be compared with the existing research. In the feature space defined in this paper, the same data set and the same algorithm as in the literature are used for comparative experiments. The algorithms used in the literature are based on the Gaussian radial support vector machine, linear kernel support vector machine, polynomial kernel support vector machine SVM and KNN algorithm.

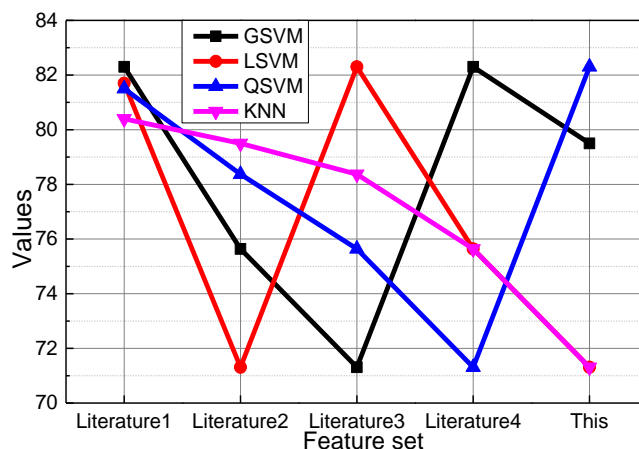


Figure 4: Experimental results of feature validity evaluation.

The experiment uses the notes dataset and adopts the method of fold cross-validation. It takes the average accuracy as the final classification performance index. The final result is shown in Figure 4, we can see that under the feature space proposed in this paper, the accuracy of the four

algorithms is higher than that in the literature. This shows that the features mentioned in this paper can improve the algorithm's ability to recognize the difficulty of the score, and it also indirectly shows the effectiveness of the features in this paper.

In the ten experiments, the difficulty matched only 8 times that the unmatched time, and the matching rate was only 66.7%. To find out the reason for the unsuccessful matching, we analyze the similarity of each candidate score in the experiment. Figure 5 lists the similarity between the template and the candidate score in each group of experiments. The red font marks the difficulty level. It can be seen from the table that although the recommendation results do not meet the requirements, the similarity of the two is indeed the highest in the difficulty level, and the similarity of candidate scores with the same difficulty level is lower. Considering that the difficulty level in the experiment comes from manual annotation, there are subjective differences which lead to the above situation. The applicability of the recommendation algorithm meets the needs of the system, and the difficulty classification will be optimized later to improve the matching rate of the recommendation system.

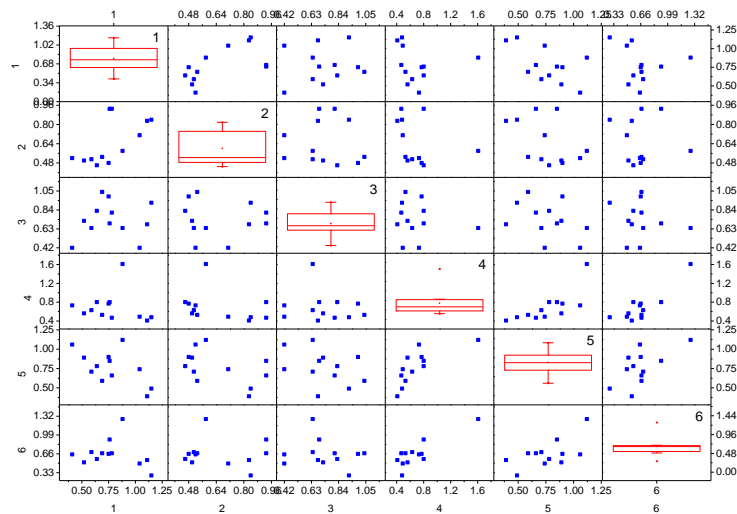


Figure 5: Candidate score similarity.

To score sight-singing, you need to calculate the similarity between the pitch sequence in the audio sequence and the template pitch sequence, and some abnormal data in the extracted pitch sequence is not caused by human factors. As showed in Figure 6, the blue represents the extracted pitch sequence, and the brown is the pitch sequence of the template score because the length of the two sequences is different. It can be seen from the figure that the segment cannot correspond through the comparison of specific data. It is found that the b segment is the singing part of the segment, which is caused by the singing error of the previous note. The difference in the sequence is unreasonable. The DTW algorithm can avoid this problem. As showing in Figure 5, a and b segments can be matched correctly.

3.2 Analysis of System Test Results

As showing in Figure 7, it is the pitch sequence curve extracted from audio, from which we can see the fluctuation of the pitch with time. However, due to the accuracy of the algorithm and the influence of noise data, there are many wild points in the pitch sequence. To better match and compare with the template, the pitch sequence needs to be smoothed. Commonly smoothing algorithms are linear smoothing and median smoothing. However, both methods require multiple

smoothing and will cause the normal pitch to shift in time, so this paper proposes a smoothing method for scene optimization.

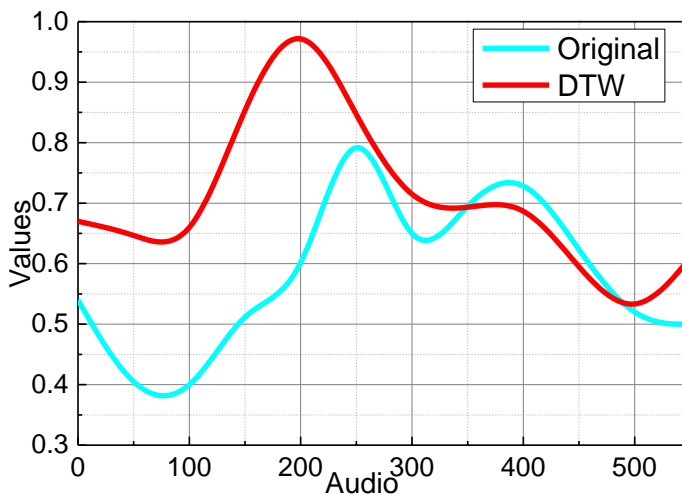


Figure 6: DTW matching diagram of the sound sequence.

Based on the algorithm for pitch extraction, the audio signal is framed. The pitch sequence we obtained is also in units of frames. Assuming that each singer is aiming at accurate singing score, then there is a wild point. It should not last long, and it should be a smooth pitch subsequence before and after it. In fact, through the analysis of the duration of the wild points in the pitch sequence of a large amount of audiovisual audio, it is found that a large number of wild points appear in the middle of the two stable signals, and last only 1 to 2 frames.

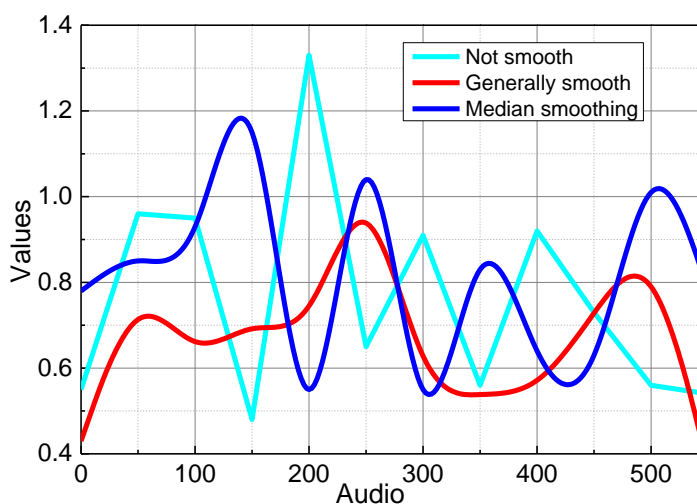


Figure 7: Effect diagram of median smoothing and optimized smoothing.

This experiment verifies the reliability of the algorithm by comparing the algorithm scoring results with the manual scoring results. The experimental data are the test questions of a certain college music exam and the results of the candidates for sight-singing WAV audio files and manual scoring. There are five judges were each audio score, overall description Figure 8.

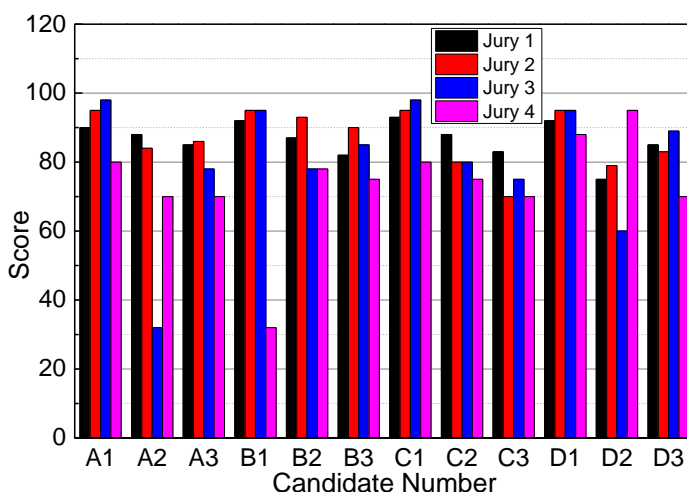


Figure 8: Teacher system test evaluation diagram.

The final experimental results are shown in Figure 9. It can be seen that although the score evaluated by the algorithm is different from the manual score. The results of the assessment algorithm is a word for algorithm evaluation of the level of sight-singing and evaluation of the results of the artificial match. We use test function of two independent sample t-test scores in both cases equal. The unequal variance t-test scores are the result of the two groups to accept the null hypothesis, significant differences in the probability of greater than 0.05. It can be considered that there is no significant difference between the results of manual scoring and algorithm scoring.

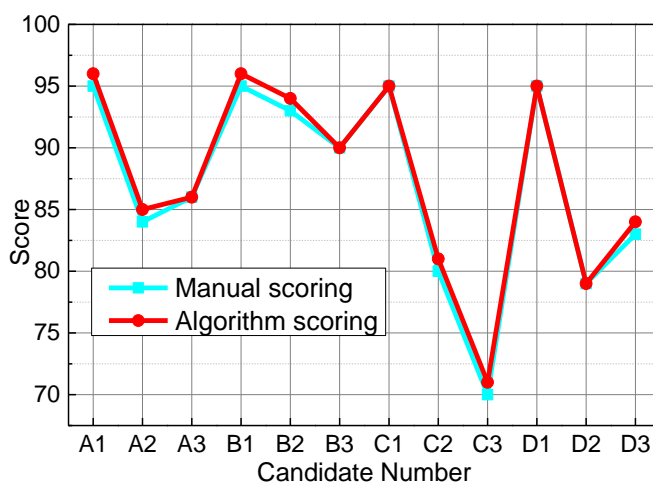


Figure 9: Manual scoring and algorithm scoring result diagram.

4 CONCLUSION

This paper provides a useful reference for the design and implementation of the personalized learning system of music visual singing by studying the model and design of the vocal music teaching multimedia computer-aided system based on voice recognition. To solve the two key problems of sight-singing ability evaluation, a score recommendation algorithm based on difficulty

features and a sight-sing scoring method based on pitch sequence matching. It was born and developed under the dual background of the rise of postmodernism in art creation and the interdisciplinary application of human-computer interaction technology and concepts. Essentially, it used sensor controllers and communication methods to capture the sound of strung instruments and more action details and transmits the data to the corresponding modules of the computer human-computer interaction platform in real-time. Electronic music conforms to the post-modernist artistic quality. It can better detect the sound of string instruments and the movement of the left and right hands and provides for the creation of the interactive electronic music modules that are needed. Based on this, it can meet the requirements of composers and string instrument players in the creation and performance of experimental interactive electronic music. The application of this achievement in the fields of music performance is also worth looking forward to it.

Meili Zhou, <https://orcid.org/0000-0003-0121-0280>

Tianzhuo Gong, <https://orcid.org/0000-0002-8045-0962>

REFERENCES

- [1] Xu, L.-D.; Xu, E.-L.; Li L.: Industry 4.0: state of the art and future trends, *International Journal of Production Research*, 56(8), 2018, 2941-2962. <https://doi.org/10.1080/00207543.2018.1444806>
- [2] Liu, L.-L.; Pang, Y.; Hu, Z.: Application of Spectrogram Analysis in Traditional Vocal Music Teaching and Multimedia Animation Vocal Music Teaching, *International Journal of Emerging Technologies in Learning*, 11(11), 2016, 64-67. <https://doi.org/10.3991/IJET.V11I11.6242>
- [3] Hinton, G.; Deng, L.; Yu, D.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, *IEEE Signal Processing Magazine*, 29(6), 2012, 82-97. <https://doi.org/10.1109/MSP.2012.2205597>
- [4] Dahl, G.-E.; Yu, D.; Deng, L.: Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 2012, 30-42. <https://doi.org/10.1109/TASL.2011.2134090>
- [5] Mohamed, A.; Dahl, G.-E.; Hinton, G.: Acoustic Modeling Using Deep Belief Networks, *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 2012, 14-22. <https://doi.org/10.1109/TASL.2011.2109382>
- [6] Netzer, Y.; Wang, T.; Coates, A.: Reading Digits in Natural Images with Unsupervised Feature Learning, *Informedness, Markedness and Correlation*, 3(5), 2011, 17-24. <https://doi.org/10.1162/NECO.2006.18.7.1527>
- [7] Dehak, N.; Kenny, P.-J.; Dehak, R.; Dumouchel, P.: Front-End Factor Analysis for Speaker Verification, *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 2011, 788-798. <https://doi.org/10.1109/TASL.2010.2064307>
- [8] Bojanowski, P.; Grave, E.; Joulin, A.: Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics*, 5(1), 2017, 135-146. https://doi.org/10.1162/TACL_A_00051
- [9] Hinton, G.; Deng, L.; Yu, D.: Deep Neural Networks for Acoustic Modeling in Speech Recognition, *IEEE Signal Processing Magazine*, 29(6), 2012, 82-91. <https://doi.org/10.1126/SCIENCE.1127647>
- [10] Powers, D.-M.: Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation, 2(1), 2011, 37-63. <https://doi.org/10.1007/978-3-642-25191-76>
- [11] Blankertz, B.; Tomioka, R.; Lemm, S.: Optimizing Spatial filters for Robust EEG Single-Trial Analysis, *IEEE Signal Processing Magazine*, 25(1), 2008, 41-56. <https://doi.org/10.1109/MSP.2008.4408441>