International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 7808-7816

https://doi.org/10.70517/ijhsa464672

The application of voice singing timbre assessment criteria in objective evaluation software algorithms

Kai Liu[1,*]
[1]School of Education Science and Music, Luoyang Institute of science Technology, Luoyang 471000, Henan, China
Corresponding authors: (e-mail: 200901501749@lit.edu.cn).

Abstract This work aims to quantify vocal singing timbre evaluation parameters that are subjective and apply them to intelligent objective evaluation software. This research first examines how vocal singing timbre is subjectively evaluated and then improves the universal assessment indices. It next looks into how to convert these standards into numerical vectors that can be used as input into an intelligent assessment system. Finally, it uses multilayer perceptrons and convolutional neural networks to model in order to extract timbre features and perform automatic evaluation. The implementation of the algorithm, including data sampling, preprocessing, embedding layer operation, intermediate layer convolutional operation, and post-processing algorithm, is also thoroughly covered in this work. The experimental findings demonstrate that the system can successfully eliminate human bias and realize timbre judgment that is somewhat consistent with subjective evaluation. Although aspects like mood and style have not yet been taken into account by the existing system, this study offers a theoretical framework and technical support for the validity and applicability of the vocal intelligent evaluation system.

Index Terms vocal singing, timbre evaluation, subjective evaluation, objective evaluation, convolutional neural network

I. Introduction

The importance of vocal art in music performance has gained a lot of attention. Tone has a direct impact on both the singer's expressiveness and the listener's aesthetic experience, making it a crucial component of vocal singing [1]. Vocal timbre evaluation has historically primarily relied on the subjective assessments of experienced judges. While this technique of review can incorporate the judges' wealth of knowledge, it also invariably introduces subjectivity and inconsistency into the evaluation standard. Thus, one of the main areas of focus in the field of vocal music research today is how to achieve the objectivity and intelligence of timbre judgment [2], [3].

Accurate and impartial timbre evaluation is crucial for identifying and developing exceptional vocal talents in vocal education and competition [4]. The conventional assessment approach has numerous drawbacks and mostly depends on the experts' subjective opinions. For instance, there could be discrepancies in the standards and comprehension of timbre across various experts, and there might even be inconsistent assessments of the same expert over time. Furthermore, the impartiality and fairness of the judgment are impacted by subjective evaluation, which is readily influenced by elements such the judges' personal feelings, degree of weariness, and psychological condition [5], [6]. Thus, there is a pressing need to design an impartial evaluation system built on clever algorithms that can successfully minimize human bias while enhancing evaluation accuracy and consistency.

We encounter numerous obstacles in our endeavor to actualize the objectivity of timbre judgment. First off, the timbre of voice singing is a multifaceted and intricate attribute that encompasses more abstract elements like resonance, vocal cavity shape, and emotional expression in addition to more fundamental acoustic characteristics like pitch, volume, and duration [7]. It is challenging to quantify these irrational emotions and turn them into numerical vectors that computers can comprehend and interpret.

Second, many high-quality labeled data points are needed as training samples for timbre evaluation. Nevertheless, gathering such information is difficult, particularly when labeling requires the assistance of qualified vocal instructors and judges, which is expensive and time-consuming [8], [9]. Furthermore, there is some individual heterogeneity in the subjective assessment of timbre, and various judges may assign different ratings to the same singing tape, adding to the difficulty of data identification.

Finally, deep learning and sophisticated machine learning algorithms are inextricably linked to the intelligence of timbre judgment. In real-world applications, these algorithms must also overcome a number of technical obstacles. For instance, there

are several topics that require in-depth research, such as how to select the best network topology and hyperparameters, handle noise and outliers in the data, and enhance the model's capacity for generalization.

With the advancement of artificial intelligence technology in recent years, an increasing number of research have started experimenting with the use of deep learning and machine learning algorithms for the autonomous evaluation of vocal timbre [10]. For instance, some studies have evaluated and classified timbre using conventional machine learning techniques like decision trees and support vector machines, and they have produced specific outcomes. Nevertheless, these techniques frequently exhibit drawbacks when handling intricate timbre characteristics and are unable to faithfully convey the minute timbre variations [11].

In fields like voice and image processing, deep learning algorithms—particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—have proven to be highly effective at extracting features [12]. Some studies have started using these sophisticated algorithms into the evaluation of vocal timbre in recent years. For instance, in some research, timbre features have been extracted from sound wave spectrograms using CNNs; in other studies, timbre changes over time have been captured using RNNs modeling continuous audio signals. Unfortunately, there are still a lot of issues with practical applications, and the majority of these investigations are still in the theoretical and experimental phases. For instance, the limited training data, high model complexity, and high computational resource consumption of these methods restrict the broad applicability of these techniques [13].

This research suggests an intelligent timbre rating system based on convolutional neural networks and multilayer perceptrons to overcome the aforementioned issues. First, we refine a set of universal evaluation indexes by thoroughly analyzing the subjective timbre evaluation criteria in vocal singing. These indications encompass more ethereal timbre characteristics like resonance and vocal cavity shape in addition to more fundamental acoustic metrics like pitch and loudness.

Secondly, we develop a fast and effective preprocessing and data sampling technique that can separate high-caliber training samples from massive volumes of audio data. Through spectrum analysis and feature extraction on audio data, we built a model framework that combines a convolutional neural network and a multilayer perceptron. To accurately capture and evaluate timbre information, the model makes full use of CNN's feature extraction capabilities and MLP's classification capability.

The experimental findings demonstrate that the intelligent timbre assessment system described in this study may successfully eliminate human bias and provide timbre evaluation that is somewhat consistent with subjective judgment. This study offers a theoretical foundation and technical support for the accuracy and applicability of the vocal intelligent evaluation system, even though elements like emotion and style have not yet been taken into account in the existing system.

II. Subjective standards for timbre assessment in vocal singing and their use

The results of the vocal singing evaluation method that is now in place may all be attributed to personal tastes and personalities. The work is universally embraced, but it also adheres to a lot of the same aesthetic principles. The growth of the voice singing industry is blatantly homogenous in terms of aesthetics [14]. As a result, a fairly uniform standard of judging has actually developed to a significant degree. This standard is recognized and influenced by the public as well as by the subjective aesthetics of the majority of assessors, who tend to converge on it. The study of the application of a single subjective criterion in objective evaluation in this research is made possible by the intrinsic homogeneity of art.

Therefore, this study's goal is very clear: before creating an objective evaluation system, it is important to clarify that the standards for both the subjective and objective evaluation modes are the same, that is, they are both grounded in the core theories of vocal music, and that they both align with the auditory standards of the common aesthetics of traditional vocal singing [15].

The primary reference indications in the vocal music evaluation method are pitch, timbre, and rhythm. These reference indices are the most readily materialized time series characteristic parameters if we think of the voice as a time series. Pitch and rhythm have been studied as part of the present objective evaluation methodologies, and software has been developed for use in some leisure activities and competitive references. Furthermore, timbre this index is less researched due to its modeling complexity and is not yet a This paper's main issue is the evaluation of vocal singing's timbre. In contrast to pitch and rhythm, timbre is a typical fuzzy evaluation index with variation, and it is more challenging to refine its physical model. This is the point at which the paper becomes challenging.

Vocal learners with normal voice conditions can typically execute works of corresponding difficulty with proper vocal training. However, each person has a distinct voice, and their training and natural vocal characteristics vary, thus they will all have somewhat varied singing tones [16]. Even following the identical training method, the ultimate sound of a voice with long, wide, and powerful vocal folds and a voice with small, narrow, and weak vocal folds may differ greatly. In terms of popular aesthetics, a strong, brilliant tone is more favored than a thin, dark tone. This is one of the guidelines taken into account when constructing the criterion for subjective evaluation developed in this paper. Second, the same person singing notes in different registers also has different sound quality. Tones in the same register are more similar. In addition to innate conditions, different laryngeal positions and different degrees of respiration produce different tonal effects. Usually, a low and stable laryngeal position produces a broader and more metallic tone, while a high and unstable laryngeal position produces a thin, narrow and

dull tone. A sound with good breath support is stable and solid, while a sound with shallow breath support is weak, wobbly and lacks stability.

The two primary facets of assessing sound quality are the two points mentioned above. In order to simulate subjective timbre perception, we will use various numerical coding techniques in this paper to express the relative merits of these two indicators. We will then use our professional experience with vocal weighting to obtain a final coding model, which we will then correspond to the node state vector of the multilayer perception network. To establish the groundwork for the final training of the perception of timbre, we must adjust the mean value and normalize the variance of the values in order to find a medium ground between good and terrible. We also need to balance the positive and negative values.

The identification problem is the primary concern and necessary stage before the evaluation in a procedure like the one described above. The identification method is a technique for using sound waveform data analysis. Software design, data modeling and algorithm study, and criterion study are the three primary components of the analytical technique. The algorithmic research component is one of the most important problems. Numerous academic studies have been conducted on recognition algorithms. Currently, advanced processing algorithms employ the artificial neural network method, which involves building an algorithm, establishing a nonlinear model based on artificial neural networks, and completing the computation of the algorithm's parameters and weights using training sequences. Prior to using artificial neural networks for sample training, the subjective evaluation criteria must first be standardized, quantified, and expressed quantitatively. The preliminary preclassification of different elements in each tone serves as the foundation for this process, which is based on subjective evaluation standards. A poor tone will have irregular overtones and a messy fundamental, big and slow amplitude, and a "bulky" or "weak downward" sound picture. On the other hand, a good tone has good overtones, a clean fundamental, tiny and quick amplitude, and a "dense" and "bright upward" sound image. Uneven overtones and a jumbled bassline with a big, slow sound amplitude are signs of poor tones, giving the impression that the sound is "bulky" or "weak and downward." Certain sounds fall somewhere in between good and bad, so it's important to identify which aspects are more "good" or more "bad," as well as how much of the good sound is composed of overtones, how much of the fundamental is composed of overtones, and how much of the good sound is composed of intangible feelings. In other words, quantitative as well as qualitative results should be provided. To put it another way, quantifiable numerical indications should be provided in addition to qualitative findings. As a result, while building the simulated neural network, quantifying it, and finishing the mathematical statement, we must consider each dimension. Network training is the process of completing the extraction of single or composite features and connecting the subjective evaluation with the quantity of network features. Different tone features correspond to different node states, and after numerous iterations of learning and training, a set of node states and feature quantities consistent with the subjective evaluation are formed. The extracted features are connected with the artificial neural network's node connection state matching.

In principle, while an individual is performing subjective evaluative tasks, different tones elicit distinct brain responses. For example, a corresponding region of the nervous system becomes active and warms up when someone hears a strong sound with disordered overtone vibrations. A sound that has a high position and consistent overtone vibrations helps you distinguish between distinct sounds and the active area of your nervous system. This implies that varying tones result in varying levels of stress and release. There are variations in the connections between neurons. The condition of these various neurons and the various types of synaptic connectivity are particular information that can be referred to in an impartial assessment method. This data is quantified using distinct algorithms that assess each sound separately, i.e., the degree to which a given sound achieves the desired levels of brightness and fullness at brief time intervals. The next steps involve assessing the sound's durability throughout the extension, identifying any color attenuation or augmentation, and determining the amount of upward and downward movement. Once you have determined your overall score for these stages, consider the quality of the two tones' transitional singing. How much the connecting process is graded before proceeding to the assessment of the lengthy phrases, and how much the timbral characteristics alter as one tone changes to another. Individual tone evaluation and long sentence evaluation are not only the sum of individual tones; they each have their own matching algorithms. As a result, additional characteristics—such as the sharpness, saturation, and overtone vibration amplitude—are required to characterize the evaluation of transition tones. The evaluation will be more accurate the more dimensions that are included. The evaluation will be more accurate the more dimensions that are included. Each individual timbre aspect, such as brightness, saturation, softness, metallicity, penetration, and other assessment angles frequently employed in subjective evaluation, can be assessed once the overall timbre has been scored. Every timbre attribute has a related vocal technique point; for instance, a high penetration grade indicates effective use of the singing. A drop in singing breathing strength is correlated with a fall in the sound penetration score in the connection of two tones. A high voice position is correlated with a high sound brightness score. A full singing cavity opening is correlated with a high sound saturation score. This phase, often known as the feedback or revision phase, can operate as the "teacher" in the vocal teaching after-school program. This "teacher" can precisely and meticulously perform the duty of guidance while excluding other states and reasons. This work primarily focuses on the early stage scoring of timbre characteristics; further research can be conducted using the feedback stage as a foundation.

Convolutional neural networks are the model utilized in the objective evaluation software program developed in this paper. Natural language processing and image processing also make extensive use of this network. These days, the findings are more

developed; they share a common location with tone feature extraction in the algorithm, but they differ in terms of parameter settings, training procedure, and other aspects. This study will then go on to describe the particular implementation procedure and potential outcomes of a more precise evaluation program.

III. Principles of objective evaluation algorithms

III. A. Data handling

Prior to building an artificial neural network, we had to gather and prepare the data. In order to achieve this, we conducted a series of data sampling procedures. We used a dataset that included recordings of multiple singers and voice students performing at varying levels of skill. We then determined the vocalists' strengths and weaknesses in terms of timbre in order to manually score the recordings and assign a numerical evaluation. This dataset includes several timbre samples together with the assessment weights that go along with them [17], [18]. A positive number would indicate the premium grade, a negative number would indicate the reverse direction quality grade, and 0 would represent the midrange timbre quality. The value should be balanced in this way. In order to finally arrange this dataset into sample training sequences, we preprocess it. A section can also be set aside as a test sequence, the scaling relationship of which can be ascertained based on the precise number of samples. Several key covariates are defined prior to preprocessing:

Record length: By adding unique markers, we can calculate the length of the given data. Anything longer than this must be trimmed, and anything shorter must be filled in. Size of timbre variety: this parameter determines the final total timbre vector dimension and the size of our timbre vectors' embedding layer.

The formula is as follows:

$$X[n] = \sum_{k=0}^{N-1} x[k] \cdot w[k] \cdot e^{-j\frac{2\pi}{N}kn} \tag{1}$$

Among them, $x[k]$is the input audio signal, $w[k]$is the window function, and N is the number of FFT points.

Algorithm order: the number of tones we want the convolution kernel to cover at once is determined by this option. An integer multiple of the algorithmic order represents the total number of convolution kernels. The level of neuron activity is controlled by the activation parameter. It is only turned on during training and turned off during testing.

Following the determination of the aforementioned parameters, preprocessing is completed in the following steps:

(1) Import a sample of data from the original data file, which is the sound recording file following interception;

(2) Data cleaning, which calls for certain signal processing in the filtering technique to handle noise, etc.;

(3) Normalization of the recordings, filling each set to the required duration and finishing the marking;

(4) Creation of an index table that associates every articulation with a number between zero and the timbre's length, converting each syllable into an integer vector.

Preprocessing gets the information ready for us. Next, we go on to the signal processing phase, which makes use of extremely specific algorithms. Model construction, the core algorithm, and the post-processing algorithm make up its three sections.

To enhance the accuracy of timbre feature extraction, an operational model must first be conceived. The primary goal of our model construction is to categorize and extract various timbre features, thereby providing the timbre evaluation value. This process is also known as the embedding layer operation. Each timbre is represented as a vector, which can correspond to single or composite features, depending on whether it has a single description or a collection of multiple descriptions. This component's job is to translate the timbre index into a representation of the timbre vector that is lower dimensional. The embedding matrix, which we obtain during the data training process, represents the table of timbre vectors that we learn from the data.

The second phase, or the intermediate layer operation, is performed to finish the overall evaluation of the vocalization of two successive tones after the above data preparation and hierarchical conceptualization. In this section, convolutional kernels of various dimensions are used in signal processing algorithms that are based on convolutional processes. As a result, after the convolution operation, each convolution kernel generates a tensor of distinct dimensions. Consequently, we must build a network layer for each convolution kernel and then combine the output of these convolutions to create a sizable feature vector.

Post-processing, which is the third phase, smoothes the findings to prevent obsessing over a certain aspect. Currently, convolutional neural networks are regularized most commonly using this step. The idea is to "disable" some neurons' capacity to fire with a specific probability. This method drives neurons to learn more beneficial qualities independently, preventing them from collectively adjusting to a single timbre feature.

We can construct a loss function to measure the mistake and control it during the procedure. This function is the optimization's minimization objective function if the entire iterative procedure is viewed as an optimization process. The cross-entropy loss function is used as the conventional loss function for the classification issue in this paper. We won't reiterate the extensive technical information here.

III. B. Hybrid recommendation algorithms

One reason for selecting the content-based recommendation algorithm is that, while it can lessen the influence of different outside factors on the algorithm itself, its use can also produce recommendation results that are easier to understand and that highlight specific traits of the matching user. Since textual data comprises the majority of the tagged items in the recommender system, the TF-IDF algorithm [7] is used to convert the textual data into feature vectors, which are then computed to determine the weights of each keyword in the sparse matrix. This process creates the token feature representation. Conversely, the item-based collaborative filtering algorithm model possesses a great generalization capacity, allowing it to filter out more complicated expression information, prevent content analysis ambiguity, and promote comparable items to the ones that users have previously enjoyed. Here, "item" refers to the individual items in the teaching platform that the recommendation algorithm may work with, such as instructors, music, videos, and other materials; they are also called "tokens" in the following contexts.

The following phases make up the majority of the voice feature-based content recommendation algorithm:

1. Representation of features: displaying in vector form the vocal characteristics that were taken from audio. Let $v_i$be the vocal feature vector and ith be the vocal sample.

2. Similarity measurement: Describe the techniques used to test how similar voice samples are to one another, including cosine similarity and Euclidean distance.

$$d\left(\mathbf{v}_i, \mathbf{v}_j\right) = \sqrt{\sum_{k=1}^{N}\left(v_{i,k} - v_{j,k}\right)^2} \tag{2}$$

$$\cos\_sim\left(\mathbf{v}_i, \mathbf{v}_j\right) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\left\|\mathbf{v}_i\right\| \left\|\mathbf{v}_j\right\|} \tag{3}$$

Among them, is the vector inner product, where $\mathbf{v}_i, \mathbf{v}_j$are vectors$\left\|\mathbf{v}_i\right\| \left\|\mathbf{v}_j\right\|$, respectively; The norm.

3. Recommendation generation: For a given vocal sample$\mathbf{v}_i$, calculate its similarity with other vocal samples in the database and select the sample with high similarity as the recommendation result

III. C. Software design

As shown in Figure 1, user-inputted vocal samples: The user provides audio files or other vocal data as vocal samples to the system. Preprocessing of the user-input vocal samples, including frame-splitting, windowing, FFT, and other operations, is done in order to extract vocal features from the data.Vocal feature extraction and representation: The vocal evaluation algorithm uses a vector representation of the voice features that were acquired after preprocessing. Vocal feature vectors are used for vocal evaluation using a content-based algorithm. Vocal evaluation may involve vocal quality scoring, similarity calculations, and other tasks.Create a report and recommendation for the vocal evaluation. Outcomes: Produce a vocal assessment report by utilizing the evaluation algorithm's results. Based on the content, suggest related vocal pieces or musical compositions.

Viewing the Evaluation Report and Recommendation Results: In order to get insight into their vocal performance and receive customized vocal recommendations, users can check the vocal evaluation report and recommendation results that have been prepared.In order to give consumers effective and precise vocal evaluation as well as tailored recommendation services, a comprehensive vocal evaluation software that integrates vocal feature extraction and content-based recommendation algorithms can be realized with the aid of such a design and flowchart.

IV. Experimental

IV. A. Vocal assessment

We essentially conducted a preliminary evaluation of voice singing timbre throughout our studies, focusing on the impact of overtones vibrating with highly distinguishable features. Some evaluation results were more in line with the subjective assessment. In general, samples with high voice position, low overtone amplitude, and strong speech penetration were evaluated higher than those with significant voice jitter and high amplitude. Students' voices and singers' voices are sampled at the same pitch and in the same language, as Figure 2 illustrates. A vocalist who adopts poorly-maintained samples or who sings at an ordinary level and is only famous for other reasons will score lower than a voice that has a high objective evaluation score, which could belong to an ordinary vocal learner instead of a famous singer. additional factors that make one famous, the ratings will likewise be poor. These are the benefits of objective assessment, and the most important thing to remember is that the software evaluation's findings should align with the findings of the vocal profession's subjective evaluation.

The results indicate that the note onset boundary's average absolute error is 26.0 ms, 80.07% of the absolute errors between the note onset boundary and the real boundary are within 50 ms, and 96.71% of the note absolute errors are fewer than 100 ms.The average absolute error ε (in semitones) and error rate er of the pitch algorithms for pitch extraction utilizing the global search and local search strategies guided by the reference sheet music are compared in Table 1. Error rate er comparison findings. When the baseline pitch-guided algorithm is used to find the fundamental cycle position in response to the reference

score, the average absolute error of the estimated pitch decreases from 0.45 semitone to 0.23 semitone, and the proportion of half-octave/octave errors decreases from 2.02% to 0.31%.

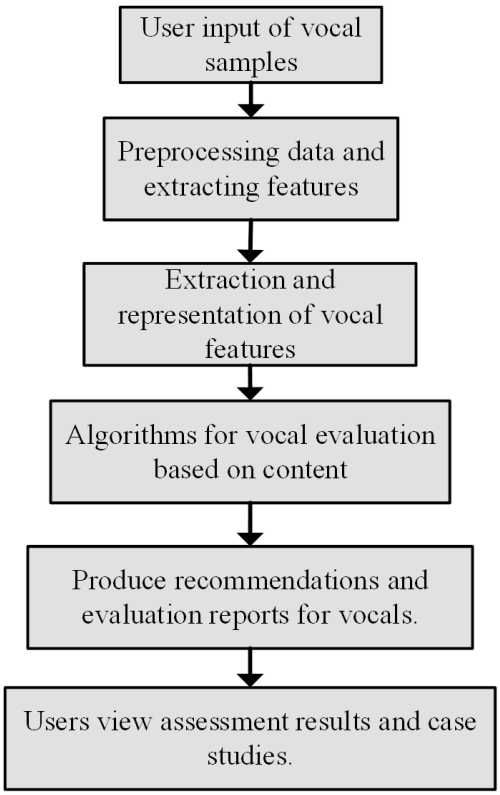

Figure 1: Algorithm flowchart in this article

Figure 2: The note onset boundary

Table 1: Pitch extraction algorithm performance

Search method	Male voice		Female voice		Comprehensive	
	ε	er	ε	er	ε	er
Global Search	0.55	3.56	0.42	1.40	0.44	2.01
local search	0.18	0.35	0.25	0.28	0.22	0.32

The automatically labeled data and the manually labeled data, respectively, were used to calculate the objective evaluation indexes proposed in Section III in order to investigate the effect of the automatic labeling algorithm's accuracy on the terminal sight-singing evaluation system's performance. The results are displayed in Table 2.

Table 2 shows that the effect on the detection of obvious pitch errors is within the tolerable error range, and the difference between the transposition amplitude calculated using the auto-annotated data and the manually annotated data is 0.302 semitones, which is less than 8% of the actual transposition amplitude. The note duration standard deviation coefficient and the tempo curve fitting residuals were calculated with automatic labeling; the errors associated with this method were constrained to 17 ms and 62.97 ms, respectively.

Table 2: Two annotations calculate the deviation of objective evaluation indicators

	Shift tone/Half tone	Difference in intervals or semitones from normal pitch	Pitch error accuracy%	Pitch error recall rate/%	Root mean square of residuals/ms for the rhythm curve	Note duration standard deviation coefficient/ms normalized
Difference between the detection findings automatic and manual annotation	0.303	0.288	78.8	85.6	17.02	63.25
Mean of test outcomes with manual annotations	3.788	0.688	-	-	91.22	228.55

The experiment's assessment of the vocalist's personal voice circumstances is erratic, and the input standard is chosen to align with popular aesthetics. The experiment's vocal singing technique and singing state, as depicted in Figure 3, will cover some of the original timbre's characteristics. In other words, focused training later in life can effectively compensate for the absence of innate voice conditions, as the resonating cavity can adjust the thin voice's innate voice and a strong respiratory support can increase the voice's penetration and fullness, leading to an extremely high objective evaluation score. As research continues, emotion and style can also be modeled, and the evaluation system will get better and better. At the moment, it only evaluates pure voice characteristics in addition to emotion; it has not yet taken into account the influence of emotion, style, or other factors.

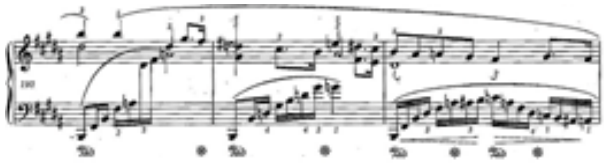

Figure 3: The note changes

IV. B. Pronunciation effect

This study looks at how well the system recognizes individual speakers, groups of speakers, and persons who are not identified. Experiments conducted on individual subjects demonstrate the system's good performance. The male speaker under study is the subject of the training database, which comprises 100 instances of 10-digit word strings, 2 strings, 3 and strings, 4. The recognition database, on the other hand, comprises the male speaker's pronunciation following a two-week break, and the library also includes 100 instances of 1 string, 2 strings, 3 strings and 4 strings. The recognition results of the system for the training and recognition sets are shown in Table 3. The length of the digit strings during the experiment is unknown.

Table 3: Outcomes of a methodical identification of particular individuals

Training set				Identification set			
Number of Sounds	String length	Error rate string number	Quantity of incorrect identifications that were corrected by addition and deletion	Number of Sounds	String length	Error rate string number	Quantity of incorrect identifications that were corrected by addition and deletion
100	1	1%/1%	0/0/1	100	1	1%/1%	0/0/1
100	2	2%/1%	0/2/0	100	2	2%/1%	0/1/1
100	3	2%/0.5%	0/2/0	100	3	3%/0.6%	0/1/2
100	4	4%/0.3%	1/1/2	100	4	4%/1%	0/2/3

The results in the table show that:

(1) This system performs exceptionally well in terms of recognition for specific persons, with a word string misrecognition rate of only 4%;

(2) When the word string length misrecognition rate rises, the majority of errors arise from deletions and additions, primarily when the numbers are read consecutively, which is also challenging for human ear discrimination. The training set for the system's multiple speaker and disrecognition studies consisted of 25 male voices pronouncing the 2- and 3-word strings ten times apiece and the 4-word string twenty times, for a total of 1,000 sounds. Pronunciation mistakes and clear endpoint detection mistakes were eliminated, leaving 952 tones available for real training.

To create the recognition set for the multi-speaker studies, we randomly picked 10 male voices for testing out of the 25 male sounds said for the training set.A total of 1,100 tones were produced by the 10 male voices, each of which uttered the 1-string

ten times, the 2-string, the 3-string, the 4-string, the 5-string, and the 6-string twenty times [19]–[21]. First, a 10-digit reference pattern library is obtained by training the system with a training set of 25 male voices pronouncing words differently. Then, using it as a priori knowledge, ten male voice pronunciations are recognized in a multi-speaker recognition set. Table 3 presents this outcome. It is uncertain how long the string was during the experiment.

Table 3 shows that: (1) the rate of string misrecognition increases progressively with string length, but the rate of numerical misrecognition does not follow this law; rather, the rate of misrecognition is lowest for 1 ∗833% when the string is the longest (6). This demonstrates that, despite an increase in string misrecognitions, the overall number of digit errors as a proportion of all misrecognitions has reduced rather than increased. Numerical misrecognition in addition and deletion errors is computed according to the accuracy of the relevant digit.

(2) Relatively strong performance is shown in terms of misrecognition rates for strings and digits. According to the experiment results, a sizable percentage of the errors are identical to those of the person-specific system and fall under the adds and deletions category. These errors primarily happen during the concatenation of digits.

In a separate experiment, the pronunciations of five more male words were recognized using the reference pattern library that was created from training the training set with the prior twenty-five male pronunciations. The male voices in this unidentified recognition set were trained to pronounce the 1-string, 2-string, 3-string, 4-string, 5-string, and 6-string 20 times apiece, for a total of 600 tones. The results of this recognition are displayed in Table 3, although the experiment's string length was not known.

V. Conclusion

The subjective evaluation criteria for vocal singing are covered in this study along with how objective evaluation software algorithms use them. In addition, this study provides a brief overview of the software algorithm's construction of a simulated neural network. The software effectively completes the extraction of voice samples' basic timbre features. The software can classify and quantitatively evaluate the overtone quality, sound saturation, brightness, penetration, and other key characteristics of various timbres through sample training and actual measurement, and the actual test verifies that it can make a more consistent evaluation of vocal timbres with the subjective evaluation standard. It is anticipated that this software will result in a useful product for the objective assessment of vocal music, which may be used in regular vocal music instruction and grow to be a significant addition to the vocal singing evaluation system.

References

[1] Cui, Y. (2023). Vocal music performance evaluation system based on neural network and its application in piano teaching. *Revista Ibérica de Sistemas e Tecnologias de Informação,* (E55), 451-464.

[2] Fernandes, J. F. T., Freitas, D., Junior, A. C., & Teixeira, J. P. (2023). Determination of harmonic parameters in pathological voices—efficient algorithm. *Applied Sciences, 13*(4), 2333.

[3] Sonkaya, Z. Z., Özturk, B., Sonkaya, R., Taskiran, E., & Karadas, Ö. (2024). Using objective speech analysis techniques for the clinical diagnosis and assessment of speech disorders in patients with multiple sclerosis. *Brain Sciences, 14*(4), 384.

[4] Wang, Z., Müller, M., Caffier, F., & Caffier, P. P. (2023). Harnessing Machine Learning in Vocal Arts Medicine: A Random Forest Application for "Fach" Classification in Opera. *Diagnostics, 13*(18), 2870.

[5] Bruder, C., Poeppel, D., & Larrouy-Maestri, P. (2024). Perceptual (but not acoustic) features predict singing voice preferences. *Scientific Reports, 14*(1), 8977.

[6] Liu, F., & Wu, J. (2023). Evaluation of Music Art Teaching Quality Based on Grey Neural Network. *Mobile Information Systems, 2023*(1), 7285914.

[7] Yuan, Y. (2024). Influencing factors and modeling methods of vocal music teaching quality supported by artificial intelligence technology. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 19*(1), 1-16.

[8] de Oliveira Florencio, V., Almeida, A. A., Balata, P., Nascimento, S., Brockmann-Bauser, M., & Lopes, L. W. (2023). Differences and reliability of linear and nonlinear acoustic measures as a function of vocal intensity in individuals with voice disorders. *Journal of Voice, 37*(5), 663-681.

[9] Gauer, J., Nagathil, A., Lentz, B., Völter, C., & Martin, R. (2023). A subjective evaluation of different music preprocessing approaches in cochlear implant listeners. *The Journal of the Acoustical Society of America, 153*(2), 1307-1318.

[10] Wenjie, B. (2024). Simulation of vocal teaching platform based on target speech extraction algorithm and cloud computing e-learning. *Entertainment Computing, 50,* 100700.

[11] Di Cesare, M. G., Perpetuini, D., Cardone, D., & Merla, A. (2024). Assessment of voice disorders using machine learning and vocal analysis of voice samples recorded through smartphones. *biomedinformatics, 4*(1), 549-565.

[12] Calà, F., Frassineti, L., Sforza, E., Onesimo, R., D'Alatri, L., Manfredi, C., ... & Zampino, G. (2023). Artificial intelligence procedure for the screening of genetic syndromes based on voice characteristics. *Bioengineering, 10*(12), 1375.

[13] Wang, Y. (2024). The effectiveness of innovative technologies to manage vocal training: The knowledge of breathing physiology and conscious control in singing. *Education and Information Technologies, 29*(6), 7303-7319.

[14] Maskeliunas, R., Damasevicius, R., Kulikajevas, A., Pribuisis, K., Ulozaite-Staniene, N., & Uloza, V. (2024). Synthesizing Lithuanian voice replacement for laryngeal cancer patients with Pareto-optimized flow-based generative synthesis network. *Applied Acoustics, 224,* 110097.

[15] Shi, Y. (2023). The use of mobile internet platforms and applications in vocal training: Synergy of technological and pedagogical solutions. *Interactive Learning Environments, 31*(6), 3780-3791.

[16] Hu, W., & Zhu, X. (2023). A real-time voice cloning system with multiple algorithms for speech quality improvement. *Plos One, 18*(4), e0283440.

[17] Li, F., Hu, Y., & Wang, L. (2023). Unsupervised Single-Channel Singing Voice Separation with Weighted Robust Principal Component Analysis Based on Gammatone Auditory Filterbank and Vocal Activity Detection. *Sensors, 23*(6), 3015.

[18] Schultz, B. G., Rojas, S., St John, M., Kefalianos, E., & Vogel, A. P. (2023). A cross-sectional study of perceptual and acoustic voice characteristics in healthy aging. *Journal of Voice, 37*(6), 969-e23.

[19] Ali, J., Jhaveri, R. H., Alswailim, M., & Roh, B. H. (2023). ESCALB: An effective slave controller allocation-based load balancing scheme for multi-domain SDN-enabled-IoT networks. *Journal of King Saud University-Computer and Information Sciences, 35*(6), 101566.

[20] Zhang, Y. (2025). An innovative deep learning method for IoT malware identification. *Mari Papel y Corrugado, 2025,* 29-37.
[21] Simon, M., & Din, S. M. (2025). Performance evaluation of self-organizing features in wireless sensor networks. *TK Techforum Journal (ThyssenKrupp Techforum)*, *2025*(1), 12–19.