Scholar Publishing Group
International Journal of Art Innovation and Development
https://doi.org/10.38007/IJAID.2022.030206
ISSN 2789-7192 Vol. 3, Issue 2: 61-78

Voice Timbre Evaluation of Broadcast Host Based on Extraction of Voice Feature Parameters

Aomar Jaine*

Madhyanchal Professional University, India

**corresponding author*

Keywords: Voice Feature Parameters, Acoustic Parameters, Broadcast Host, Voice Timbre Evaluation

Abstract: The continuous development of science and technology has continuously improved people's quality of life. People's definition of health is also clearer, and more attention is paid to laryngeal diseases and the quality of voice. In real life, accurate evaluation of voice quality and timbre is not only beneficial to the diagnosis and treatment of laryngeal diseases. At the same time, it also plays a vital role in the selection and training of broadcasting and hosting talents. The current evaluation methods for related professional voices have shortcomings such as time-consuming, labor-intensive and highly subjective. In order to overcome these defects and explore how to help the training of broadcast hosts, this paper conducts in-depth research on the processing method of speech signals and the objective evaluation method of artistic voice timbre from the perspective of signal processing. Through experimental analysis, it is found that the objective evaluation method of voice timbre based on acoustic parameters F0, F1, F3 can better realize the objective evaluation of the voice timbre of the broadcast host. Among them, the accuracy rate of the evaluation method based on multiple feature parameter extraction reaches 89.2%.

1. Introduction

Voice is an important tool for human language expression. It plays a very important role in frequent social interactions. Timbre is the unique characteristic of sound. Its quality directly affects a person's ability to express language, and even his future. Excellent voice timbre conditions are the basis for the success of voice workers. They pay more attention to the voice than ordinary people. The quality of the voice and timbre directly affects their professional life. However, when selecting and cultivating talents, the evaluation of vocal talents is basically that the voice experts use their own experience to listen and feel the timbre of the test subjects as a whole, and use simple scores or

comments to represent the evaluation results. Due to the differences in the evaluation standards among the evaluators and the influence of various factors, the evaluation results are lacking in objectivity, accuracy and fairness to a certain extent. Moreover, the results of subjective listening perception evaluation often cannot fully reflect the vocal condition of vocal talents, and it is difficult to provide an effective reference for various problems in vocal music teaching.

In this paper, the voice signal of the broadcast host staff is analyzed by using the knowledge of computer and acoustics. At the same time, it made an objective evaluation of their voice timbre according to various factors and laws of their pronunciation. Finally, it quantifies the pros and cons of sound from an acoustic point of view. It explores how to obtain higher-quality sound from the source of sound. This method of using computer to objectively evaluate professional voice timbre has the advantages of scientific, visual and non-invasive. Moreover, it can also overcome the shortcomings of traditional subjective evaluation methods, such as large workload, time-consuming and labor-intensive, and low flexibility. Therefore, it can ensure the objectivity of talent selection to the greatest extent.

The innovation of this paper is to combine the evaluation of voice timbre with computer and acoustic technology. It also offers unique insights into voice care issues for broadcasters. This move is bound to create a fair and objective environment for the selection and training of broadcasting and hosting practitioners, and at the same time, it can detect voice problems in a timely manner and prevent minor problems.

2. Related Work

With the rapid development of science and technology, more and more technologies are gradually moving towards integration to achieve cross-border development. In recent years, methods based on speech feature parameter extraction have emerged in an endless stream, and researches on the voice timbre of broadcast hosts are also abound.

You M pointed out a cough detection method based on pattern recognition technology, which aims to objectively assess the quantity and intensity of coughs. From the spectrum of cough, he saw that the energy spectrum of cough signal is widely distributed over the entire frequency band, which is very different from speech signal. To find the difference between cough and other audio, he used non-negative matrix factorization to extract the spectral structure of the two from the signal, and designed a filter bank for feature extraction. This makes it more suitable for cough detection than manually designed filter banks [1].

Subba Ramaiah V proposed a speaker classification system. It uses Tangent Weighted Mel Frequency Cepstral Coefficients (TMFCCs) as characteristic parameters. A Lion algorithm is also used to cluster audio streams detected by voice activity into specific speaker groups. Finally he utilized low-energy frames as well as high-energy frames with stronger effects, improving the performance of the proposed system [2].

Macasero R pointed out that automatic speaker recognition models are the basis for building various speaker representations, pattern analysis and engineering models. In this process, the accuracy of classifier recognition is the most difficult part of system design, and it is especially important for models that are deployed in real-time scenarios. At the same time, he also proposed a new deep learning speech recognition model, which is designed to automatically recognize speech words [3].

Zaydn N aimed to study how to capture and identify the acoustic features of dysphonia patients. So he compared acoustic speech parameters obtained from the PRAAT software with parameters for

automatic feature extraction in dysphonia patients. As a result, he found that the jitter of acoustic parameters was significantly different between normal and dysphonic patients. Based on this, he designed an acoustic recognition system for patients with dysphonia [4].

Yermolenkina L aimed to study how speech and language communication alter vocal characteristics. Based on this, he conducted a service evaluation of male participants receiving speech and communication group therapy. It also had a 12-month follow-up to investigate its vocal characteristics and level of verbal communication. On this basis, he proposed a program to promote voice change and skill development, aiming to cultivate people's phonological skills and phonological self-awareness [5].

Schweinberger S R pointed out that the research of Mobile Ad Hoc Network based on Named Data Network (NDN based MANET) has made new progress. At the same time, he observed that most of the NDN-based MANET research will choose to use the all-wireless broadcast method for dynamic routing as their solution in the MANET environment [6].

Sun Y pointed out that voice hygiene is an important part of the voice training process. So we should pay attention to voice hygiene to maintain a healthy voice. He assessed the vocal hygiene of middle school students in the 7th and 8th grades of voice change and analyzed the differences between the various vocal health test points. At the same time, he proposed that before starting the test, each question should be checked by experts in the relevant field for the validity of the test, and to ensure that the tested students can understand the relevant questions [7].

3. Signal Preprocessing and Speech Feature Parameter Extraction

3.1. Voice Signal Preprocessing

The key to the objective evaluation of the host's voice lies in the extraction of acoustic parameters. No matter what method is used to extract and analyze parameters, the voice signal must be processed in advance, including digitization, pre-emphasis and framing [8]. Preprocessing is a key step in analyzing voice signals and extracting acoustic parameters, and it is also the premise for successful voice feature analysis. Next, the article will start with signal digitization, pre-emphasis and framing, and gradually explore the preprocessing process of voice signal.

3.1.1. Signal Digitization

Before any kind of sound enters the computer, it is just an external voice signal, that is, an analog signal. Therefore, to process and analyze the sound, it is necessary to first convert the analog signal into a digital signal [9]. In this paper, some voice signals of broadcast hosts are recorded through microphone and computer sound card, and the useful and digitized voice data files are directly obtained for subsequent use.

In general, converting an analog signal into a digital signal requires three stages: sampling, quantization, and encoding. Next, it will be explained one by one. The so-called sampling is the random information capture of the analog signal sequence, but the random here does not mean arbitrary, but refers to the same time interval. Sampling is to reduce the storage and processing of a large amount of data, and it is the discretization of the original analog information. Quantization refers to the realization of the fluctuation of the analog signal into a certain value to achieve digital expression. Quantization is carried out on the basis of sampling, so the characteristics of the sampled data will be intuitively presented in the quantized values. After the quantized value is obtained, the next step is the final encoding. It all knows that computers use binary, but humans are

used to decimal. So in order for a computer to be able to process data and information, it must convert the previously quantized values to binary [10]. In fact, the process of converting to binary is just a simple encoding process. In real life, people often split binary values into binary or multi-valued digital signal streams in order to allow digital signals to propagate through various media. However, the specific conversion principle will not be repeated here.

Since analog signals can be converted to digital signals, presumably digital signals can also be converted to analog signals. In fact, analog signals and digital signals can be converted to each other. The analog signal is generally quantized into a digital signal by a PCM pulse code modulation (Pulse Code Modulation) method, that is, different amplitudes of the analog signal correspond to different binary values respectively. In practice, for the storage and transmission of signals, people often use 24-bit or 30-bit encoding. The digital signal is generally converted into an analog signal by the method of phase-shifting the carrier. Binary digital signals are used in computers, computer local area networks and metropolitan area networks. In the 21st century, both binary digital signals and analog signals converted from digital signals are actually transmitted in the computer wide area network. However, it is still the digital signal that has more application and development prospects [11].

3.1.2. Pre-emphasis and Framing

After acquiring the voice signal data file of the broadcast host, pre-emphasizing and framing the signal can make the processed signal better meet the actual needs. At the same time, it is of great significance to improve the accuracy of subsequent data processing.

In the process of processing the corresponding signal files, it finds that some of the voice signals appear in the voice signal with a relatively low frequency. On the other hand, the frequency of the other part of the voice signal is relatively high. In order to ensure the integrity of the speech signal to the greatest extent, it is necessary to emphasize the corresponding part of the speech to ensure the original characteristics of the speech signal, and the process of emphasizing the relevant part is called pre-emphasis. The purpose of pre-emphasis is to emphasize the high-frequency part of the speech to remove the influence of non-critical factors in the digitization process and increase the high-frequency resolution of the speech.

The pre-emphasis of the speech signal provides a guarantee for our subsequent processing. However, the volume of the voice data files after the aggravation processing is often very large. Therefore, in order to reduce the burden of computer processing and improve the data processing capability, it is necessary to segment the speech signal [12]. Segmentation processing is a common method for computers to process general data, which is manifested in speech signals by dividing the signal into frames. A schematic diagram of the framing of the speech signal is shown in Figure 1.

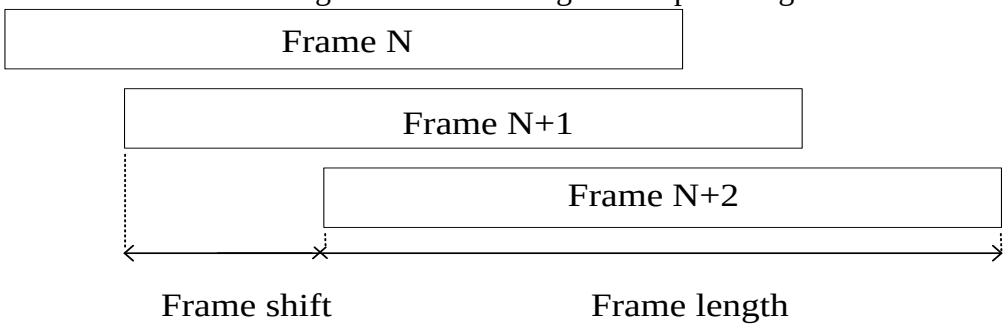


Figure 1. Framing diagram of speech signal

3.2. Speech Recognition

After a series of preprocessing, a relatively complete speech data file can be obtained [13]. But for machines to understand what humans say and perform relevant actions based on human voice commands, it requires giving machines the ability to read and understand speech. In the process of continuous development, people have developed a speech recognition technology, which can truly realize the language interaction between humans and machines. Research on speech recognition often covers many disciplines, and its application fields are also very wide. Figure 2 shows the main application areas of speech recognition.

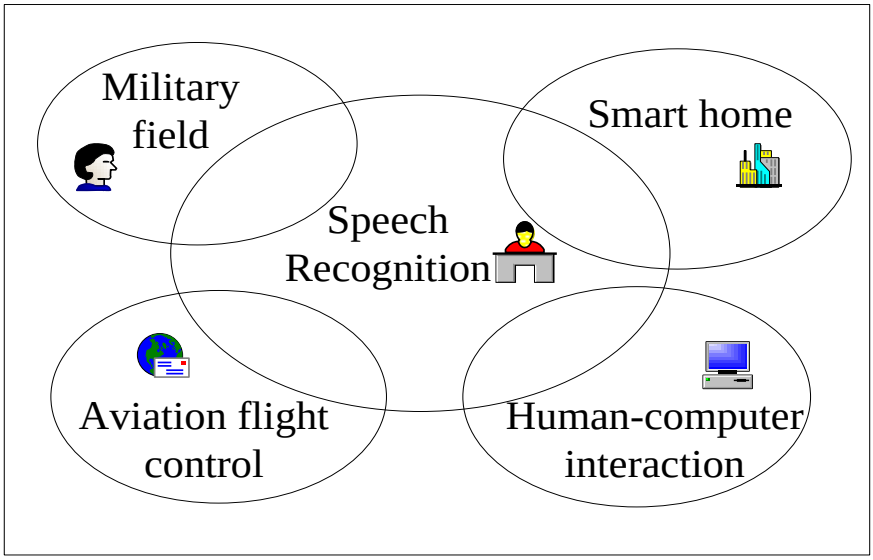


Figure 2. Speech recognition application fields

Speech recognition is also divided into different categories according to the size of the vocabulary, the way of pronunciation, and the objects to be recognized. The following is its basic classification.

(1) According to the size of the vocabulary, it can be divided into: a small vocabulary recognition system covering dozens of words, a medium vocabulary recognition system covering hundreds of words, and a large vocabulary database recognition system covering thousands to tens of thousands of words [14].

(2) According to the pronunciation method, it can be divided into: isolated word speech recognition system that requires appropriate pauses for input words, connective word speech recognition system that the input words are all pronounced accurately, and continuous speech recognition system that the input speech is natural and fluent.

(3) According to the recognition object, it can be divided into: a specific person speech recognition system that can recognize a person, a non-specific person speech system that can recognize any person, and a multi-person recognition system that can recognize a certain type of person.

3.2.1. Speech Feature Parameter Extraction

The selection of speech feature parameters and extraction algorithms plays a crucial role in the entire speech recognition system [15]. Preprocessing, speech enhancement, endpoint detection and

other links are all to pave the way for feature parameter extraction, which is convenient for subsequent training and identification. In this process, the process of extracting the feature parameters containing semantic information components in the speech through an algorithm is called speech feature extraction.

But before parameter extraction, it is also necessary to understand the characteristics of the voice and the generation of the speech signal. The vocal cords are an important foundation for human vocalization. At the same time, the generation of voice signals also requires the participation and completion of multiple parties of the vibration organ and the nervous system. Figure 3 is a time domain model of speech signal generation. As can be seen from the figure, the generation of speech is the result of the joint action of multiple layers, and various disturbances from the outside world will be encountered in the process of its generation. Since the production of voice is complex and diverse, the appearance of voice is also multi-functional and multi-structured. Therefore, the evaluation of the voice must also be multi-parameter, and the timbre of the broadcast host's voice must be evaluated comprehensively and objectively.

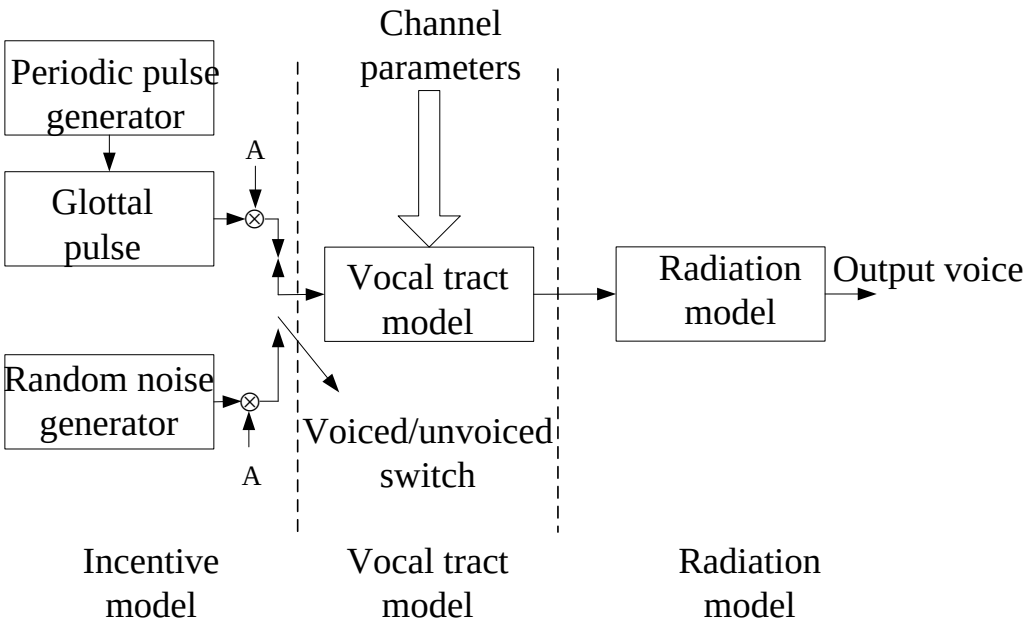


Figure 3. Time domain model of speech signal generation

3.2.2. Acoustic Parameter Selection

Different acoustic parameter selections will affect the final feature parameter extraction [16]. Features are the outstanding performance of things that are different from other things, and are the key to distinguishing things. Therefore, when it wants to classify or identify things, it actually extracts different characteristics of things, and finally judges things through the performance of the characteristics. There are many kinds of characteristic parameters of speech. According to the physical characteristics of speech, the characteristic parameters can be divided into the following categories:

(1) Feature parameters based on vocal tract model

In the research process of practical things, people regard the sound channel as a system composed of multiple tubes with different cross-sectional areas connected in series. In this system, the most common acoustic tube model parameter is the linear prediction coefficient.

(2) Feature parameters based on auditory perception characteristics

The characteristic parameter based on auditory perception is a parameter that simulates the human ear's ability to perceive sound according to the inhomogeneity of the human ear's hearing perception frequency range. Therefore, these parameters are more in line with the perceptual characteristics of acoustics. The most common feature parameters are Linear Prediction Cepstral Coefficients (LPC) and Mel Frequency Cepstral Coefficients (MFCC).

3.2.3. Fundamental Frequency Extraction

Pitch is one of the three major properties of sound [17]. Except for some extreme cases, the pitch is determined by the fundamental frequency of the sound. Therefore, the terms pitch and fundamental frequency are often used interchangeably in the literature. Fundamental frequency extraction has a wide range of applications in sound processing. Its most direct application is to identify the melody of music. But it can also be used for speech processing, such as aiding speech recognition in tonal languages and recognizing emotion in speech. However, since the fundamental frequency of the sound often changes with time, the fundamental frequency extraction usually divides the signal into frames and then extracts the fundamental frequency frame by frame. The method of extracting the fundamental frequency of a frame of sound can be roughly divided into the time domain method and the frequency domain method. The spectrum will have peaks at integer multiples of the fundamental frequency. The basic principle of the frequency domain method is to find the greatest common divisor of these peak frequencies. The basic frequency extraction process of the speech signal is shown in Figure 4.

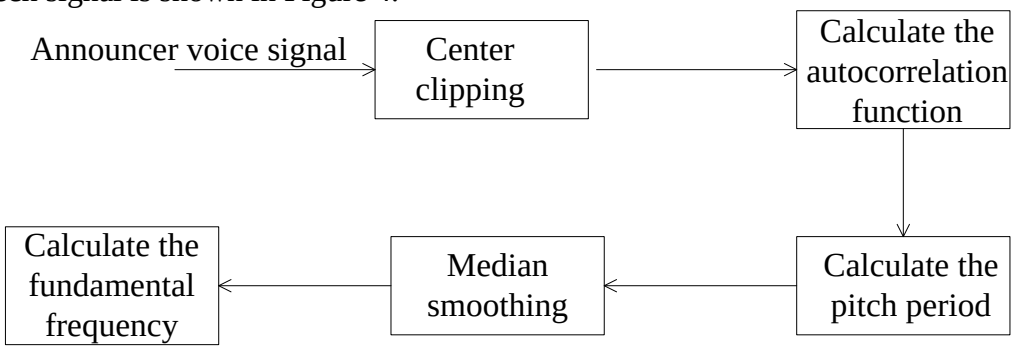


Figure 4. The process of extracting the fundamental frequency of the voice signal

3.3. Voice Timbre Evaluation Technology of Broadcasting Host

Timbre refers to the distinctive characteristics of different sounds in terms of waveforms [18]. Therefore, timbre can be simply understood as the characteristic of sound. But timbre is not like other physical properties of sound, which can be clearly linked to corresponding physical laws, such as pitch and vibration frequency, such as speed and time and displacement, such as color and wavelength. Timbre does not have a prescribed physical law. Like how people look, it is not determined by a single factor. The combination of different facial parts forms a whole, a style, which is presented in our brains, and so is the tone.

Timbre is a combination of a large number of different factors [19]. But these combinations are integrated by our brains into a single model for easy perception, so it will naturally use the word timbre to describe it. However, in the professional field, the selection of talents cannot be achieved only through simple perception and feeling. Therefore, in order to unify the standards and realize

the planning and standardization of talent selection, people have developed a voice condition evaluation system with five levels, as shown in Table 1.

Table 1. Evaluation of voice condition

Score	Sound quality	Objective distortion description
5	Very good	Imperceptible
4	Okay	Perceptible, acceptable
3	Generally	Perceptible, slightly disgusting
2	Difference	Obnoxious ,but bearable
1	Very bad	Intolerable

The table divides the difficult-to-describe voice conditions into five levels, realizing the abstraction into the concrete. However, in the actual operation process, it is found that only manual evaluation cannot achieve batch and reproducibility. Therefore, speech recognition technology was introduced [20]. In the field of broadcasting and hosting, the commonly used speech recognition technologies are: Hidden Markov Model and Neural Network.

3.3.1. Hidden Markov Model

Hidden Markov models are different from regression and classification models that deal with independent sample data. It is used to process time series data, that is, data with time series relationships between samples [21]. In this regard, it is very similar to the Kalman filter algorithm. In fact, the algorithms of Hidden Markov Model and Kalman Filter are essentially the same. It's just that the hidden Markov model assumes that the hidden variables are discrete, while the Kalman filter assumes that the hidden variables are continuous. Hidden variables are one of the key concepts in Hidden Markov Models. People can directly understand it as a variable that cannot be directly observed, that is, the meaning of the word Hidden in the hidden Markov model. In contrast to hidden variables are observed variables, that is, variables that can be directly observed. The ability of the hidden Markov model is that it can estimate what the corresponding hidden variable sequence is based on the given sequence of observed variables, and make predictions about future observed variables. The sequential hidden Markov model is shown in Figure 5.

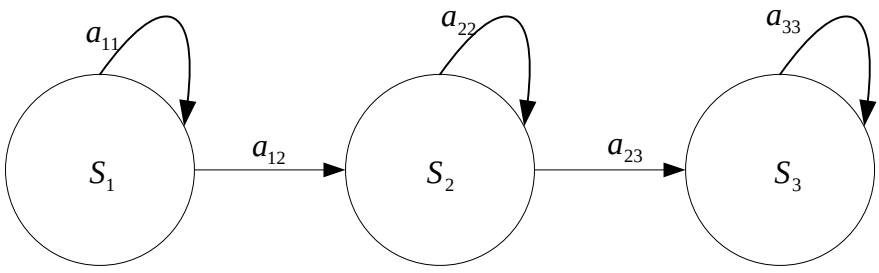


Figure 5. Sequential hidden markov model

3.3.2. Neural Networks

After talking about the hidden Markov model, let's take a look at the performance of the neural network. When it comes to this, people's first reaction is the nerve cells in the organism's brain. It is

reported that there are 14 billion neuron cells in the human brain, and its complexity is no less than the structure of network nodes. The artificial neural network is actually a set of biomimetic models established by people imitating biological neurons. In order to generalize it to other fields, the neural system is mathematicalized and closely connected with the field of artificial intelligence, resulting in artificial neural networks. Artificial neural network has unparalleled advantages and characteristics. First, the ability to learn independently. Artificial neural network is a biological neural network that simulates autonomously, so its first feature is self-learning ability. Organisms can process information hundreds of millions of times through neurons, and in the process continue to evolve themselves and put them into artificial neural networks. Computers are also able to learn and train autonomously, continuously improving their processing efficiency in the process.

Second, Lenovo storage capacity. The continuous development of artificial nerves has led to other neural networks. Among them, the feedback neural network can use association to store data and information. In this way, in the face of huge data information, the network can realize storage and calling through the association between data, which not only saves the storage space but also effectively improves the calling speed of the data. In this storage process, the neural network can also continuously optimize and upgrade through autonomous learning, so as to find the optimal storage solution for each stored process. Figure. 6 is a neural network structure diagram.

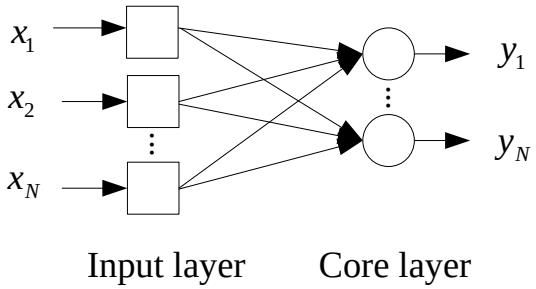


Figure 6. Neural network structure

4. Voice Timbre Evaluation Based on Extraction of Various Feature Parameters

Traditional speech feature extraction often uses neural network technology. It has multiple layers of input and output, so it can fit infinitely the functions of each output layer. If it assumes that the speech signal is v, then input it into the neural network at this time, and it can get a vector of length L. The final output function of the neural network looks like this:

$$M(n) = [v_L^1, v_L^2, ..., v_L^L] \quad (1)$$

Among them, v_L^1 represents the output result of the speech signal in the first layer.

People are used to making predictions before a neural network has come up with a final function. The predicted results are then compared with the results generated by the neural network to test the reliability of the method. The expected output function of the neural network is:

$$e(x) = [e_1, e_2, ..., e_n] \quad (2)$$

Among them, n is the number of iterations of the neural network. The error signal for the n-th iteration is defined as:

$$D(m) = D_L(n) - y_L(n) \quad (3)$$

$$y_L(n) = \frac{1}{L/2}\sum_{n=1}^{N/2}(D_n - D_i) \quad (4)$$

Among them, N is the weighted sum of errors after n iterations. If the final value is smaller, it indicates that the actual error of the neural network is smaller. The larger the value, the larger the actual error. Then the error energy is defined as:

$$E(n) = \frac{1}{3}\sum_{l=1}^{L} D_L^2(n) \quad (5)$$

In order to analyze the timbre of the voice, it is first necessary to introduce a sound capture function, which is used to identify the language signal and convert it into a machine-recognized digital signal.

$$\lambda_{ij}{}^2 = (\left|median(w_{ij})\right| / 0.6654)^2 \quad (6)$$

Among them, *median* represents the voice recognition module, which can realize the signal conversion of voice.

Once the sound is captured and transformed, it can then use a neural network to train and analyze it, and finally get the desired result.

$$\sigma_{ij}^2 = \frac{1}{2}\sum_{i=1}^{N}(\lambda_{ij} - \tilde{\lambda})^2 \quad (7)$$

Among them, σ represents the process of processing the audio digital signal. N is the length of the sound signal.

However, the results of the sound information processed by the neural network are often relatively simple, and cannot provide data support for our subsequent analysis of different timbre conditions. Therefore, based on the neural network, it integrates a variety of feature parameters and redesigns the method based on speech feature parameter extraction. In this method, K is the characteristic parameter, and its value is as follows:

$$K = \sum_{m=1}^{N-m} x_i(m) y_i(m+k) \quad (8)$$

$$Z = K_i(K+T) \quad (9)$$

$$R_n(k) = \frac{1}{N}\left|\sum_{m=0}^{N-m+1} x_m\right| \quad (10)$$

$$R_i(k) = \sum_{i=0}^{N-i+1}\left\|x_i\right| - x_i(m+k)\right| \quad (11)$$

For the same voice information, it sets four different characteristic parameters for it. Therefore, the characteristics of the voice information can be found from different levels and angles. Among them, x and y are the position information of the sound respectively, and m is the breadth of the sound. T is the period of the sound, and Z describes the distribution and period of the sound. R depicts the emotional characteristics of the sound, and it has two parameters set, which

correspond to different situations.

But in the process of feature parameter extraction, some other factors, such as information interference, often bring challenges to our extraction. Therefore, in order to reduce the interference from the external environment, the sound signal is first emphasized.

$$\varpi_{ij} = \mathrm{sgn}(\varpi_{ij})(\left|\varpi_{ij}\right| - N) \quad (12)$$

$$T = (\sigma_{nj}^2 / \sigma_{nj}) \cdot \sqrt{2\ln(M \log_2 N)} \cdot (1 + 1/\ln(j+1)) \quad (13)$$

$$\sigma_{ij} = \mathrm{sgn}(\varpi_{ij}) \frac{\varpi_{ij}^4}{\delta^3} \quad (14)$$

Among them, ϖ is the result after the information is emphasized, and N is the other unimportant branches in the signal. T represents the periodic processing of the signal, and σ_{ij} represents the processed valid information.

After getting valid information, it will be put into the neural network for training in batches.

$$S(k) = \sum_{i=2}^{4} L_j \sigma_i(k) \quad (15)$$

$$s = \sum_{k=1}^{n} a_k^i s_i(n-k) \quad (16)$$

$$A = 1 - \sum_{k=1}^{n} a_k^i S^{-k} \quad (17)$$

Among them, S represents the training process of eigenvalues, a represents the effective information, and σ represents the information processed by the neural network.

In order to test and verify the final results, it also optimizes the traditional prediction method and introduces a new verification technology. Among them, Q is the final result of the verification, and its value will be jointly affected by the information integrity e and the information characteristic value n.

$$Q = \sum_{n=1}^{N-m} e_i(n) e_{i+1}(n+1) \quad (18)$$

The information integrity model is described as follows:

$$y = \frac{\partial e(n)}{\partial w_{ij}(n)} \quad (19)$$

In this information integrity model, there is a certain quantitative relationship between i and n. Its relationship is shown in formula (21). Therefore, their values will directly affect the final data integrity, which is also a guarantee for the data.

$$W = \sum_{i=0}^{n} C_n^i \geq k \quad (20)$$

$$C_n^i = 0, \quad i > n \quad (21)$$

In the above formula, C represents the data relationship between data and speech feature values, and W describes the sum function of the data relationship. Based on this, it can use deep learning and feature parameter extraction to obtain basic feature information about voice timbre. The following article will analyze and study its related characteristics through specific experiments.

Before the start of the experiment, a statistical analysis was carried out on the voices of the students in different broadcast hosts. The statistical results are shown in Figure 7.

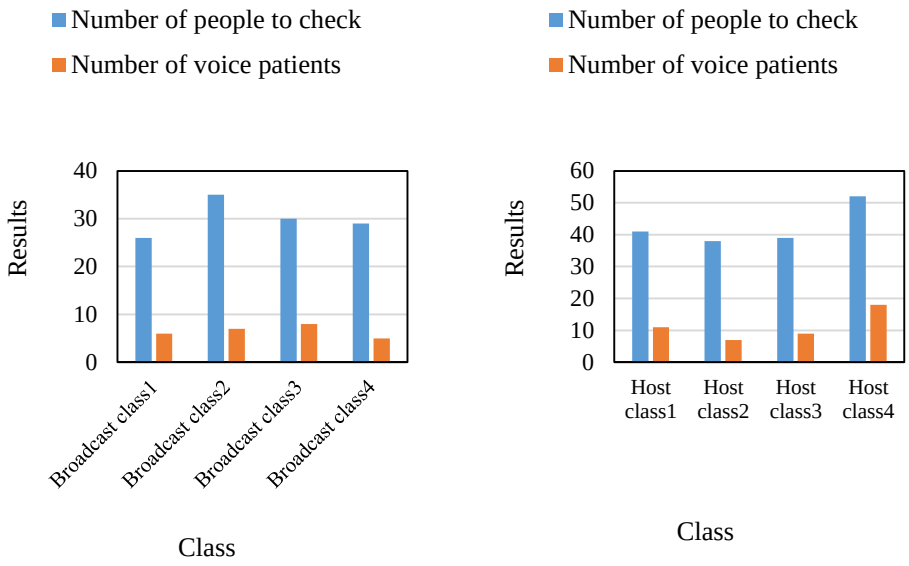


Figure 7. Analysis of voice situation of students majoring in broadcasting and hosting

It can be seen from the figure that there are different degrees of voice problems for the students in the broadcasting class and the host class. Among them, the highest number of people with hidden diseases is more than 20, which is nearly one-third of the class. And in the statistics, it is also found that some students have no understanding of their own voice problems. This poses a great risk to their throat health.

A person's vocal tract is variable, and each person's vocal tract characteristics are different. Therefore, the voice signal sent out also has the characteristics of multiple deformation and irregularity, which makes the extraction of characteristic parameters quite complicated, and also brings difficulties to the accurate detection. In order to extract voice features as accurately as possible, it selects different algorithms and makes statistics on the recognition rates under different signal-to-noise ratios. The results are shown in Table 2.

Table 2. Recognition rates of different feature extraction methods under different signal-to-noise ratios

	No noise(%)	30dB(%)	20dB(%)	10dB(%)	Average value(%)
LPCC	87.6	85.1	80.95	78.2	82.08
MFCC	88.5	86.8	82.5	76.5	83.59
Ours	96.7	92.9	86.6	84.2	89.04

Table 2 shows that in the absence of noise interference, the three algorithms can better identify the speech signal, and the algorithm recognition rate is as high as 96.7%. However, when the signal-to-noise ratio continues to improve, the recognition rates of the three methods are all declining. In particular, when the signal-to-noise ratio reaches 10, the lowest recognition rate has dropped to 76.5%.

After understanding the performance of different methods under different signal-to-noise ratios, the three algorithms were optimized and upgraded based on deep learning. The training time and recognition time are shown in Table 3.

Table 3. Training, matching and recognition times for different methods

	LPCC	MFCC	Ours
Training time(s)	10.7	11.4	11.2
Match time(s)	1.4	1.7	1.3
Recognition time(s)	12.2	12.9	9.9

Table 3 shows that after the continuous strengthening of deep learning, the algorithm matching time and recognition time are further shortened, and the matching time is only 1.3 seconds.

Under the training of deep learning, the recognition accuracy and matching time of the algorithm have been greatly improved. But in order to test its application in the voice timbre evaluation of the actual broadcast host, the following experiments are designed. Among them, Table 4 is the acoustic parameters of 18 broadcasting and hosting students and the results of their subjective evaluations by experts.

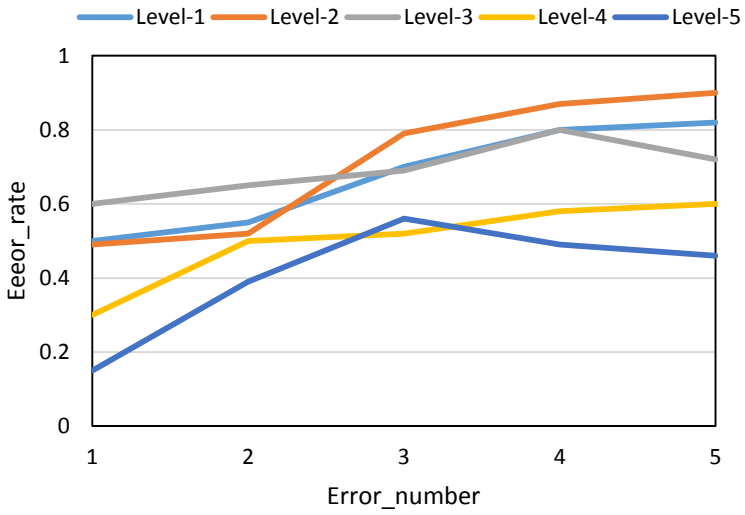
Table 4. Acoustic parameters and subjective evaluation results

serial number	gender	F1/HZ	F3/HZ	F0/HZ	Subjective evaluation results
1	Female	578.5	2147.6	516.2007	1
2	Female	621.9	2440.8	518.3610	1
3	Female	453.1	2139	486.3678	2
4	Female	404.2	2567.4	527.3618	5
5	Female	558.6	2909	478.2578	4
6	Female	458.1	2909	598.2356	5
7	Female	632.6	2909	358.2178	3
8	Female	408.6	2909	260.2891	5
9	Female	398.6	2909	2708.2698	1
10	Male	354.9	2518.4	226.6648	2
11	Male	297.4	2654.3	268.7824	1
12	Male	335.8	2592.3	164.25112	3
13	Male	291.2	2418.2	255.8139	4
14	Male	386.1	2558.7	308.9557	2
15	Male	338.4	3016.3	268.0468	5
16	Male	402.8	2409.8	251.4853	1
17	Male	368.8	300.7	247.7823	3
18	Male	424.6	326.4	270.0462	2

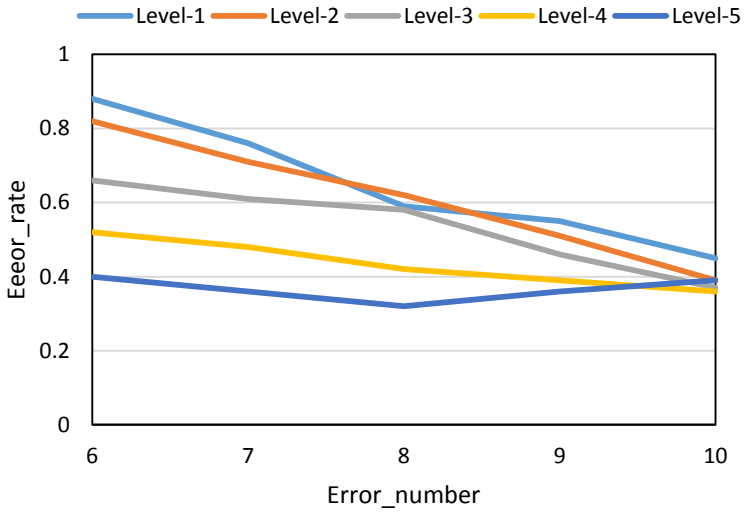
Table 4 shows that, according to the grading table (Table 1), it can be known that the rank 5 is excellent, and the rank 1 is poor. Among the 18 students, the expert evaluations gave a scale ranging

from 1-5.

For the integrity of the data, it chose to import the data of all 18 students into the algorithm. Figure 8 shows the distribution of different voice levels and evaluation error rates.



A



B

Figure 8. Relationship between voice level and evaluation error rate

Figure 8A shows that during the first five experiments, the error rate of evaluating students with higher grades, that is, students with a grade of 5, showed a certain upward trend with the increase of the number of experiments, reaching a maximum of 0.5. Figure 8B shows that in the subsequent experiment process, the evaluation error rates of students of different grades are decreasing, which shows that the accuracy of the algorithm has reached a balance after many experiments.

However, the above picture only analyzes the accuracy and recognition of the algorithm from the perspective of level. Next, it will analyze the recognition and evaluation of the algorithm under different parameters from the perspective of characteristic parameters. Figure 9 is the emotion

recognition degree under different parameter features.

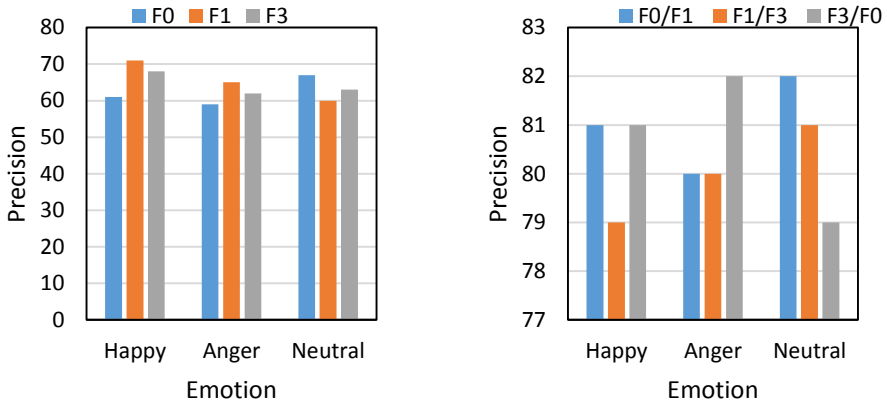


Figure 9. Emotion recognition under different parameter features

Figure 9 shows that under a single parameter matching, the algorithm's recognition degree for different emotions is basically about 60%, which can basically meet the requirements. However, under the fusion of various parameters, the algorithm's ability to recognize different emotions reaches more than 80%. Based on this, it can basically prove the effectiveness of our algorithm. Figure 10 is a performance comparison of different feature parameter extraction methods.

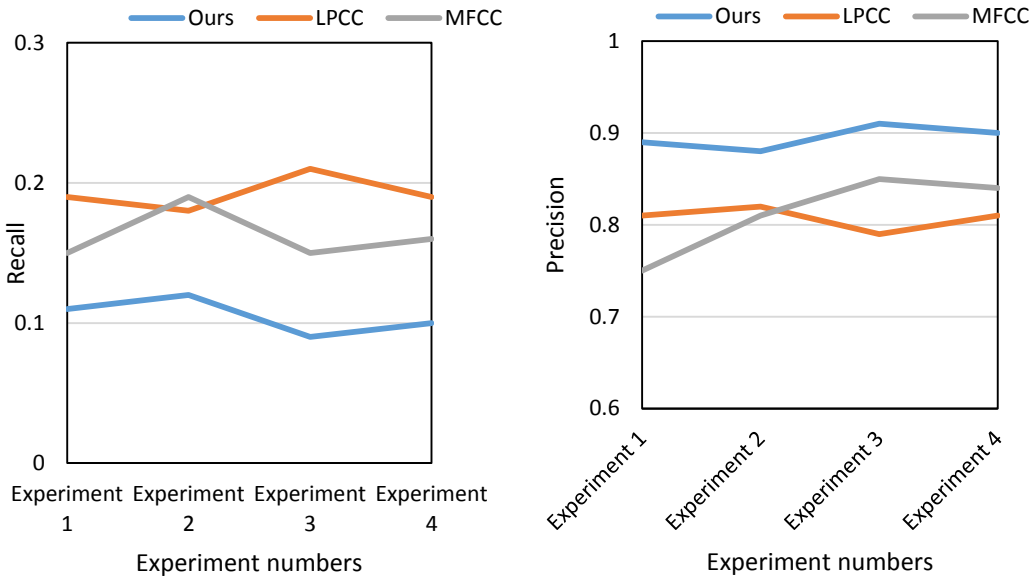


Figure 10. Performance of different feature extraction methods

Figure 10 shows that the algorithm has a low recall rate during the experiment, which basically remains around 0.1. And in terms of accuracy, the algorithm is also technically superior, and the comprehensive accuracy can reach 0.9.

Because of the uniqueness of the voice of the broadcast host, it selects different parameters to conduct a comprehensive study on it. The evaluation accuracy of different methods for voice timbre is shown in Table 5.

Table 5. Accuracy of evaluation of voice timbre by different methods

Preferences	F1	F2	F3	F0/F1	F1/F3	F0/F3	F1/F3F0
Ours	59.3	48.1	74.1	63.0	77.8	81.5	89.2
LVQ	37.4	37.04	48.2	25.9	81.5	77.8	81.9
MFCC	43.1	40.3	56.7	49.8	70.3	73.9	82.6

Table 5 shows that, compared with a single feature parameter evaluation method, the evaluation of multiple feature parameters tends to have a higher accuracy. Among them, the accuracy rate of the method based on the fusion of three feature parameters can reach 89.2%, far exceeding the other two methods.

5. Discussion

Because of its special professional needs, broadcasting and hosting professionals tend to pay more attention to the conditions of voice and timbre in the selection of related talents. The method based on multiple feature parameter extraction can provide a fair and objective evaluation for each practitioner. On this basis, the ability test and evaluation of relevant practitioners can also timely discover the defects of the broadcast host and avoid broadcast accidents in the industry. At the same time, it is also possible for broadcasting and hosting practitioners to adjust their vocal habits in time to ensure the health of the larynx.

6. Conclusion

This paper starts with sound processing, and firstly introduces some speech processing techniques in acoustics. Then it focuses on the methods and patterns of speech recognition. Finally, it proposes a method based on multiple feature parameter extraction, and successfully applies it to the voice timbre evaluation of broadcast hosts. However, in the practice process of using multi-feature parameter extraction to evaluate the voice timbre of the broadcast host, the paper does not put forward the solution and corresponding improvement measures pertinently. At the same time, the research on multi-feature parameter fusion is not very in-depth. In the future, the article will continue to deepen the research on multi-feature parameter fusion, and focus on the methods and measures, aiming to provide suggestions for the adjustment and protection of the voice tone of the broadcast host.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] You M , Wang H , Liu Z. Novel feature extraction method for cough detection using NMF. Iet Signal Processing, 2017, 11(5):515-520. https://doi.org/10.1049/iet-spr.2016.0341

[2] Subba Ramaiah V , Rajeswara Rao R . A novel approach for speaker diarization system using TMFCC parameterization and Lion optimization. Journal of Central South University, 2017, 24(11):2649-2663. https://doi.org/10.1007/s11771-017-3678-3

[3] Macasero R . Remembering Rex Cornelio: Months on from the murder of Philippines radio host Rex Cornelio we speak to those who knew him about his bravery and his awful death. Index on Censorship, 2020, 49(4):18-22. https://doi.org/10.1177/0306422020981251

[4] Zaydn N , Doanyiit S . Voice Hygiene Evaluation at 7th and 8th Grade Secondary School Students at Voice Change Period via Vocal Health Test. Universal Journal of Educational Research, 2018, 6(8):1771-1776. https://doi.org/10.13189/ujer.2018.060821

[5] Yermolenkina L I . Communicative Strategies Interaction As A Discursive Mechanism Of A Convergence Radio Host's Image Forming. Tomsk State Pedagogical University Bulletin, 2019(4):150-154.

[6] Schweinberger S R , Eiff C , Kirchen L. The Role of Stimulus Type and Social Signal for Voice Perception in Cochlear Implant Users: Response to the Letter by Meister et al. Journal of Speech Language and Hearing Research, 2020, 63(12):1-2.

[7] Sun Y , Song C . Emotional speech feature extraction and optimization of phase space reconstruction. Xi'an Dianzi Keji Daxue Xuebao/Journal of Xidian University, 2017, 44(6):162-168.

[8] Bai J , Shi Y , Xue P. CFCC feature extraction for fusion of the power-law nonlinear function and spectral subtraction. Xi'an Dianzi Keji Daxue Xuebao/Journal of Xidian University, 2019, 46(1):86-92.

[9] Nguyen T S , Li Z , Su G. Hydro-mechanical behavior of an argillaceous limestone considered as a potential host formation for radioactive waste disposal. Journal of Rock Mechanics and Geotechnical Engineering, 2018, 10(6):59-77. https://doi.org/10.1016/j.jrmge.2018.03.010

[10] Jermsittiparsert K , Abdurrahman A , Siriattakul P. Pattern recognition and features selection for speech emotion recognition model using deep learning. International Journal of Speech Technology, 2020, 23(4):1-8. https://doi.org/10.1007/s10772-020-09690-2

[11] Shamila S , Snekhalatha U , D Balakrishnan. Spectral Analysis and Feature Extraction of Speech Signal in Dysphonia patients. International Journal of Pure and Applied Mathematics, 2017, 113(11):151-160.

[12] Farahani G . Robust Feature Extraction Using Autocorrelation Domain for Noisy Speech Recognition. Signal & Image Processing An International Journal, 2017, 8(1):23-44. https://doi.org/10.5121/sipij.2017.8103

[13] Haider F , Garcia S , Luz S . An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(2):272-281. https://doi.org/10.1109/JSTSP.2019.2955022

[14] Wang C , Tao L , Ding Y. An adversarial model for electromechanical actuator fault diagnosis under nonideal data conditions. Neural Computing and Applications, 2022, 34(8):5883-5904. https://doi.org/10.1007/s00521-021-06732-x

[15] Hallek M , Boukamcha H , Mtibaa A , et al. Dynamic programming with adaptive and self-adjusting penalty for real-time accurate stereo matching. Journal of Real-Time Image

Processing, 2021, 19(2):233-245. https://doi.org/10.1007/s11554-021-01180-1

[16] Pal D , Singhal R , Bandyopadhyay A K . Parametric Optimization of Complementary Split-Ring Resonator Dimensions for Planar Antenna Size Miniaturization. Wireless Personal Communications, 2021, 123(2):1897-1911.

[17] Kumar S , Agrawal R . A comprehensive survey on meta-heuristic-based energy minimization routing techniques for wireless sensor network: classification and challenges. The Journal of Supercomputing, 2021, 78(5):6612-6663.

[18] Zhao Z , Feng Z , Liu J. Stand parameter extraction based on video point cloud data. Journal of Forestry Research, 2021, 32(4):1553-1565.

[19] Mills M , Stoneham G , Davies S . Toward a Protocol for Transmasculine Voice: A Service Evaluation of the Voice and Communication Therapy Group Program, Including Long-Term Follow-Up for Trans Men at the London Gender Identity Clinic. Transgender Health, 2019, 4(1):143-151. https://doi.org/10.1089/trgh.2019.0011

[20] Farkhana M , Hanan A A , Suhaidi H. Energy conservation of content routing through wireless broadcast control in NDN based MANET: A review. Journal of Network & Computer Applications, 2019, 131(4):109-132. https://doi.org/10.1016/j.jnca.2019.01.004

[21] Miraghaei H , Best P N . The nuclear properties and extended morphologies of powerful radio galaxies: the roles of host galaxy and environment. Monthly Notices of the Royal Astronomical Society, 2017,1(4):1-18. https://doi.org/10.1093/mnras/stx007