Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7086599, 17 pages
https://doi.org/10.1155/2022/7086599

*Research Article*

# Objective Evaluation Method of Broadcasting Vocal Timbre Based on Feature Selection

**Chaohui Lv**,[1,2] **Hua Lan**,[1] **Ying Yu**,[1] **and Shengnan Li**[1]

[1]*School of Information and Communication Engineering, Communication University of China, Beijing 100024, China*
[2]*Key Laboratory of Acoustic Visual Technology and Intelligent Control System, Ministry of Culture and Tourism, Beijing 100024, China*

Correspondence should be addressed to Chaohui Lv; llvch@cuc.edu.cn

Received 18 September 2021; Revised 9 January 2022; Accepted 7 May 2022; Published 26 May 2022

Academic Editor: Ghufran Ahmed



Broadcasting voice is used to convey ideas and emotions. In the selection process of broadcasting and hosting professionals, the vocal timbre is an important index. The subjective evaluation method is widely used, but the selection results have certain subjectivity and uncertainty. In this paper, an objective evaluation method of broadcasting vocal timbre is proposed. Firstly, the broadcasting vocal timbre database is constructed on Chinese phonetic characteristics. Then, the timbre feature selection strategy is presented based on human vocal mechanism, and the broadcast timbre characteristics are divided into three categories, which include source parameters, vocal tract parameters, and human hearing parameters. Finally, the three models of hidden Markov model (HMM), Gaussian Mixture Model-General Background Model (GMM-UBM), and long short-term memory (LSTM) are exploited to evaluate the timbre of the broadcast by extracting timbre features and four timbre feature combinations. The experiments show that the selection of timbre features is scientific and effective. Moreover, the accuracy of the LSTM network using the deep learning algorithm in the objective evaluation of the broadcast timbre is better than the traditional HMM and GMM-UBM, and the proposed method can achieve about 95% accuracy rate in our database.

## 1. Introduction

Speech recognition and synthesis have made breakthroughs in recent years, and more and more attention has been paid to the speech timbre research. However, the objective evaluation of broadcasting vocal timbre belongs to the specialist areas. The vocal timbre database can be divided into three categories, which include reading language, fluent language, and spontaneous language. Some studies have shown that broadcasting knowledge can be used in the database construction. New rhythmic features are put forward to the analysis of relevant discussions in broadcasting, such as vocal expression features, discourse stress, parataxis, and compound rhythmic phrases, and they are integrated into rhythm and text annotation specifications in database [1]. At present, PRAAT software is commonly used to analyze recorded audio with the aid of a computer [2]. In addition, the relationship between vocal timbre and emotional perception in words is also researched, and the results show vocal timbre can convey the emotional meaning [3].

From the results of speech feature extractions, the evolution of audio features can be divided into time-domain, frequency-domain, time-frequency-domain joint features and depth features. Since the end of 1950s, time-domain features have played an important role in audio analyses and classifications. Frequency-domain features appeared in order to analyze the frequency spectrum of audio signals, such as tone and formant. By the end of 1960s, joint time-frequency feature extraction algorithm appeared. With the development of deep learning, depth features began to be widely used in audio signal processing algorithms [4]. Since 2010, depth features have been applied to speech enhancement, separating speech and music audio signals, and automated audio classification [5–7].

From the perspective of the speech recognition model, the application of speech signal can be roughly divided into three categories, including vocal print recognition, speech recognition, and emotion recognition [8]. The classifiers for speech recognition tasks include traditional classifiers and deep learning algorithms, involving HMM, Gaussian Mixture Model (GMM), support vector machine (SVM), and extreme learning machine (ELM) [9–11]. At present, the role of acoustic parameters is analyzed in the objective evaluation of artistic vocal, and methods of it are proposed based on error back propagation (BP) and learning vector quantization (LVQ) [12]. Piano sound is also studied on keys influencing and its quality evaluation. And a double neural network composed of Hopfield and general regression neural network (GRNN) is implemented to evaluate the piano audio. Signals are processed and weight ratios of the output are calculated by the least square method [13].

Figure 1 shows the basic framework of objective evaluation of vocal timbre proposed in this paper. The main three steps are as follows:

(1) Construction of broadcasting vocal timbre database: this paper focuses on selections of vowels and auxiliary sounds, voiceless, and voiced, as well as words and sentences which can reflect the difference from the broadcasting vocal timbre and daily spoken. Software PRAAT is used to mark the phonetic layer, phonological layer, and prosodic layer

(2) Feature optimization based on the selection strategy of timbre features considering the mechanism of human vocalization: the selected feature parameters are required to reflect the human vocal mechanism as far as possible. Therefore, human vocal timbre features are classified by the feature selection module combining principal component analysis (PCA), and three feature categories are obtained including source parameters, vocal tract parameters, and human hearing parameters. To remove the correlations between features, PCA is exploited to reduce the dimension of the above three feature categories, and finally, 74-dimensional timbre feature parameters are obtained [14, 15]

(3) Timbre grading by the objective evaluation method: HMM [16], GMM-UBM, LSTM, and other models are used to compare different objective evaluation methods [17–20]

The main contributions of this paper are three parts. (1) Based on the broadcasting knowledge and Chinese phonetic characteristics, a database of broadcasting vocal timbre was constructed and annotated. (2) Combining the mechanism of human vocalization and PCA technology, we screened out the characteristic parameters of broadcast timbre and divided them into three categories which include source parameters, vocal tract parameters, and human hearing parameters. (3) We conducted fusion and nonfusion experiments on the three types of timbre features, respectively. Among them, we used four timbre feature combinations for fusion and conducted comparative research on objective evaluation of broadcast timbre on models such as HMM, GMM-UBM, and LSTM.

## 2. Related Work

Language is an important communication tool between people, and artistic language pays more attention to the expressiveness of language. Announcers employ the vocal organs to send out voices with personal characteristics, and listeners get different perceptions through the auditory organs and then achieve the purpose of conveying thoughts and spreading emotions. Traditional broadcast vocal timbre evaluation relies on manual evaluation, which is susceptible to the personal preferences, professional qualities, and life experience of the participants, and has greater subjectivity and uncertainty. In this context, this paper is devoted to develop a set of scientific and complete objective evaluation system of broadcasting vocal timbre.

The construction of the speech database has been booming in recent years, and various speech databases for different research needs have appeared. Due to different environment and background noises, Bang et al. directly captured broadcast data from broadcast channels, including news, current affairs, culture, documentaries, dramas, children, and entertainment, and manually constructed an English phonetic database of 7 types of evaluation data [21]. He et al. built CUCBNC, a Chinese language broadcast news speech database. By analyzing the related discourses of broadcasting, they proposed new prosodic features, including voice expression features, text stress, paraphrase groups, compound prosodic phrases, and integrated these prosodic features into the database [1]. Not the same as research in [1, 21], the data in this paper came from a clean environment, there was only one speaker with a clear voice, and the noise was small. In addition, the voice data was based on the theoretical knowledge of broadcasting and human body acoustics, focusing on selecting vowels, consonants, unvoiced, and voiced sounds that can reflect the tone of the broadcast voice, and the three-level evaluation was conducted by judges with professional qualities in the field of broadcast hosting and used PRAAT software to mark the pinyin layer, phonological layer, and prosodic layer of the speech data.

In the field of speech, different features are selected according to different research tasks. Al-Qaderi et al. used prosodic features and short-term spectral features, combined with GMM and SVM to achieve speech recognition in human-computer interaction [22]. Al-Kaltakchi et al. utilized four feature combinations based on Mel frequency cepstral coefficient (MFCC) and power normalized cepstral coefficient (PNCC), as well as seven fusion methods, combined with *i*-vector and ELM to identify the speaker [23]. The above research can find that it is not that the more features are selected, the higher the accuracy rate is. The characteristic parameters selected in this paper required that the characterization of timbre characteristics was as complete as possible and could accurately reflect the

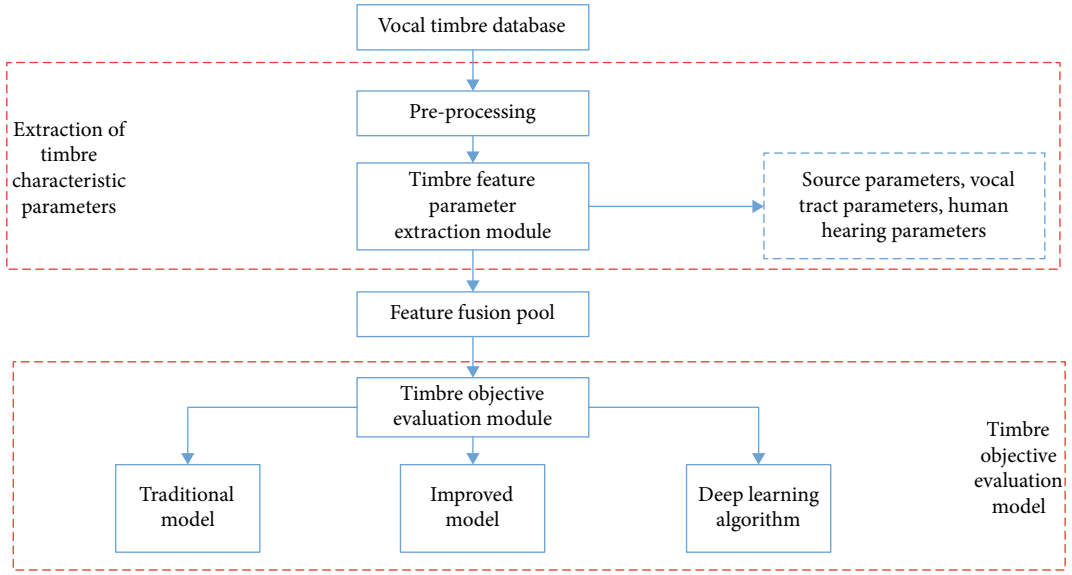


Figure 1: Basic framework of objective evaluation of vocal timbre.

mechanism of human vocalization. Therefore, from dozens of pairs of characteristic parameters, PCA and the principle of human vocalization were used for features selection, and the source parameters, vocal tract parameters, and human hearing parameters were obtained after dimensionality reduction.

In addition to extracting discriminative features from the speech signal, the choice of the classification model is also crucial for the speech recognition system. Al-Kaltakchi et al. proposed the simultaneous use of PNCC and MFCC, together with acoustic modeling using a GMM-UBM, and realized speech recognition under different signal to noise ratios [24]. Furthermore, deep learning networks have been widely developed in the field of speech recognition. Jahangir et al. employed a deep neural network (DNN), taking MFCC and time domain features as input to build a speech recognition model. DNN obtained better classification results compared with five machine learning algorithms that were recently utilized in speaker recognition [25]. Same as the speech recognition system, in the process of constructing the objective evaluation system of the broadcast timbre, it is not only necessary to extract the features of the timbre characteristics but also to select an appropriate classification model. The broadcast speech signal is a complex time-varying signal. When inputting the speech signal into the traditional shallow acoustic model, only a few levels of linear or nonlinear transformation are usually carried out. In this paper, in addition to the typical HMM and GMM-UBM models, the LSTM model is also utilized as an objective evaluation model. LSTM is a variant of recurrent neural network (RNN). By adding a gating mechanism to control the information transmission in the recurrent neural network, it can handle complex time-varying signals and has stronger modeling capabilities for the nonlinear dynamic process of human speech.

## 3. Construction of the Broadcasting Vocal Database

The speech samples were selected from the raw data of Communication University of China (CUC). These speech data were collected during preliminary and reexamination of broadcasting and hosting major from 2012 to 2016. The material format is WMV, and the recording environment is professional recording studio. Besides Adobe software system, the hardware equipment mainly includes computer Mac Pro for data acquisition and processing, Sony ecm-nv1 super directional condenser microphone. The sampling frequency is 44.1 KHz, and the quantization bits are 16 bit. In the selected 6000 speech samples, the speaker ratio of male and female students is 1 : 1, and short articles and newsletters were selected covering political hot spots, science, and technology culture. The speakers read in the same environment. Finally, teachers made a three-level evaluation on the thickness of vocal timbre to form the initial broadcasting vocal timbre database. And all the teachers were faculty of the school of broadcasting and hosting arts of CUC.

The initial vocal timbre database was further processed to remove the uncertainty influence and improve the classification accuracy. The construction of the broadcasting vocal timbre database is mainly divided into the following five steps:

(1) Format conversion: use ProCoder software to extract audio information from original video files and save them as lossless compressed wave audio file format

(2) Middle section interception: middle section of each wave file was intercepted by MATLAB program. And each audio file was saved as 20 seconds to meet the actual requirements of human hearing in subjective evaluation

(3) Noise removal: if the speech sample is processed by noise reduction in the later stage, it will cause distortion, and part of the timbre characteristics will be lost in the feature extraction process. Therefore, the human ear recognition method is used to remove the speech samples with noises generated by irrelevant factors

(4) Phoneme coverage improvement: referring to " Mandarin and Broadcasting Pronunciation," the speech samples are selected including all vowels and vocal tones as far as possible by counting the coverage of each phoneme in speech samples

(5) Pitch marking: the phonological layer and the prosodic layer of the speech samples are labeled by PRAAT software. In addition, it is necessary for professionals to confirm the label for each selected speech sample. Thin and thickness were selected to focus on each pair of evaluation terms. Other evaluation terms, such as brightness, tightness, and other evaluation indicators, need only a small number of tags from evaluators to reconfirm and replace in the later stage. Table 1 lists the subjective evaluation terms and objective parameters of the vocal quality broadcasted by electronic equipment, which also has a strong reference significance for the subjective evaluation of broadcasting vocal timbre

Through the above steps of the sample selection and processing, a dataset is completed with unified format, replaceable tag, wave format, and 20 seconds time length. Each file size is 1.68 M with 44.1 KHz sampling frequency, and the quantization bit is 16 bit. According to the thickness of timbre, the dataset is divided into three levels, *A*, *B*, and *C*. Each level contains 100 speech samples of male and female students, and the total dataset has 600 samples. Figure 2 shows the constitution of the broadcasting vocal timbre database.

## 4. Objective Evaluation Models of Broadcast Vocal Timbre

*4.1. HMM.* The hidden Markov model is used to describe a Markov process with hidden unknown parameters. It is a double random process without memory. Memoryless means that the system is in the current state, and the future state has nothing to do with the past state. One of the double random processes refers to the implicit random process that uses a Markov chain with a finite number of states to simulate changes in the statistical characteristics of the voice signal, and the other refers to the random process of the observation sequence associated with each state of the Markov chain. In the process of normal human pronunciation, such a double random process will occur, because the speech signal can be regarded as an observable sequence, and the other unobservable sequence is the parameters of the words, sentences, and segments generated in the brain. The HMM model can just simulate such a process. On the one hand, it can characterize speech signals with different characteristic parameters; on the other hand, it can also characterize the conversion process between signals well.

When HMM is applied to the objective evaluation system of timbre, the following three problems need to be solved, and corresponding algorithms are proposed for each problem:

(1) Valuation problem: given model parameter $\lambda$ and observation sequence $O$, calculate the maximum probability $P(O|\lambda)$ of the observation sequence $O$ under model $\lambda$. To solve this problem, we need to use a forward-backward algorithm to calculate the probability of each HMM producing a given observation sequence $O$ and then evaluate the optimal HMM model

(2) Decoding problem: given model parameters $\lambda$ and observation sequence $O$, find the optimal hidden state sequence. To solve this problem, the Viterbi algorithm needs to be used to decode the hidden state sequence. The Viterbi algorithm is a general decoding algorithm based on dynamic programming and the method of finding the shortest path of the sequence

(3) Learning problems: given the observation sequence $O$, the model parameters are unknown, how to maximize the output probability of the observation sequence by adjusting the model parameters. This paper utilizes the maximum expectation (EM) algorithm to solve; using the recursive idea, the local maximum value is obtained after continuous iterative optimization model parameters

*4.2. GMM-UBM.* GMM can approximate the distribution of any data by weighting multiple Gaussian probability density functions. But in the speech recognition training process, each target speaker needs to be modeled, and then the GMM models of all the trained speakers are found and tested. GMM has a strong characterization ability for actual data, but the stronger the characterization ability is, the larger the parameter scale is, and the more data is needed to drive to obtain a more generalized general model. GMM-UBM introduces many nontarget speakers' voice and uses the probability model of the spatial distribution of these nontarget speakers' voice features to train a GMM. This GMM does not have the identity of a specific target speaker. It is a priori model of the target speaker model, and then the target speaker's speech parameters are fine-tuned on this model. With the advantage of UBM, only a small amount of correction is needed when training each target speaker model individually. The model obtained in this way not only greatly improves the recognition rate but also reduces the amount of calculation.

Before training GMM-UBM, we need to initialize the model and set the initial parameters of the model. For the determination of parameter $\lambda = \{\mu, \delta, \alpha\}$, the $K$-means algorithm is usually used to roughly estimate the range of the parameter. $\alpha$ is the weighting coefficient of the model. $\mu$ is the expected value of the model. $\delta$ is the variance of the

Table 1: Subjective evaluation terms and objective parameters of vocal quality.

| Serial number | Subjective evaluation terms | Description of objective parameters |
|---|---|---|
| 1 | Wide | Advantages of wide frequency band, small distortion, large dynamic range, uniform frequency distribution, and outstanding energy in medium and low frequency band |
| 2 | Narrow | Narrow frequency band, lack of high and low frequency, medium frequency too prominent, short reverberation time |
| 3 | Bright | Moderate medium and low frequency, enough high frequency energy, small distortion, and rich harmonic decay process |
| 4 | Dark | Lack of medium and high frequency, 5-6 KHz has obvious attenuation and short reverberation time |
| 5 | Thin | Lack of medium and low frequency, 300-500 Hz has obvious attenuation, and the sound energy is insufficient |
| 6 | Thick | Medium and low frequency energy is strong, low frequency is full, high frequency is moderate, distortion is small, and reverberation is moderate |
| 7 | Soft | Low and medium frequency response range is wide, and distortion is small |
| 8 | Hard | Lack of low frequency, medium and high frequency, and high frequency harmonic decay too fast |
| 9 | Thorough | Advantages of small distortion, good transient response, enough medium and high frequency reverberation, wide and uniform frequency response |
| 10 | Paste | Adequate low frequencies, long reverberation time, and lack of medium and high frequencies |
| 11 | Pine | High frequency, low frequency, good transient response |
| 12 | Real | The medium and low frequency energy is enough, the high frequency is moderate, the distortion is small, and the loudness is high |
| 13 | Crude | Low frequency energy is large, and medium and high frequency energy is small |
| 14 | Fine | Lack of low frequency, moderate medium and high frequency, insufficient reverberation, and low loudness |
| 15 | Tip | Lack of low frequency, medium and high frequency, high frequency energy is too large, and distortion is large |
| 16 | Circular | Wide frequency band, minimal distortion, good transient response, and moderate reverberation |
| 17 | Hair | Adequate high frequency and medium high frequency components, and the distortion is large |
| 18 | Sand | The passband distortion is large and accompanied by transient distortion |
| 19 | Stuffy | Lack of medium and high frequency, serious attenuation above 3-4 KHz, and excessive low frequency energy |
| 20 | Brittle | Lack of low frequency, medium and high frequency partial, frequency response is uneven |
| 21 | Clear | Wide and uniform frequency response, rich medium and high frequency, moderate reverberation |
| 22 | Muddy | Lack of medium and high frequency, excessive reverberation, and poor transient response |
| 23 | Plump | Rich in medium and low frequencies, moderate in high frequency, and appropriate in loudness |
| 24 | Shriveled | Lack of medium and low frequency reverberation, reverberation aftersound is too short |
| 25 | Clean | Small distortion, good transient response, and moderate reverberation |
| 26 | Three-dimensional | Wide frequency response, small distortion, moderate reverberation, and large dynamic range |

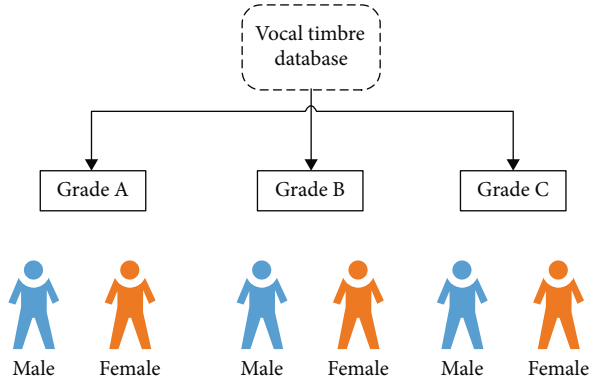


Figure 2: Broadcasting vocal timbre database.

model. After initializing the parameters, the EM algorithm is used to iteratively generate local optimal model parameters. According to the model's initial parameter $\lambda$, we train to estimate the new model parameter $\bar{\lambda}$, so that $P(X|\bar{\lambda}) \geq P(X|\lambda)$. Then, train and estimate with the new model parameter $\bar{\lambda}$ until the model meets the convergence condition.

After using the EM algorithm to train and estimate the UBM model that uses the vocal features of the nontarget speaker as background data, it is only necessary to adaptively match the trained model parameters to the GMM model of the target speaker to match the target speaker's model. The parameters are fine-tuned on the original UBM model without the need to perform GMM modeling for each speaker. The Gaussian distribution obtained by

training is fitted to the target speaker's data by using the maximum posterior probability (MAP) algorithm. The offset probability relative to the $i$-th Gaussian distribution obtained by UBM training is

$$P(i|x_i) = \frac{\alpha_i p_i(x_t)}{\sum_{j=1}^{M} \alpha_j p_j(x_t)}. \quad (1)$$

The characteristic parameter sequence of the target speaker is $X = \{x_1, x_2, \cdots, x_N\}$, and $M$ is the order of the model.

After the target speaker model is obtained through adaptive training, the method of calculating the log-likelihood ratio is used to complete the recognition task by calculating the probability score of the speech to be tested on the GMM and UBM models. The log-likelihood ratio calculation equation can be defined as

$$S(X) = \log \frac{P(X|\lambda_{\text{GMM}})}{P(X|\lambda_{\text{UBM}})} = \log P(X|\lambda_{\text{GMM}}) - \log P(X|\lambda_{\text{UBM}}), \quad (2)$$

where $\lambda_{\text{GMM}}$ is the target speaker GMM model parameter, and $\lambda_{\text{UBM}}$ is the target speaker UBM model parameter.

*4.3. LSTM.* The speech signal is a complex time-varying signal. When the speech signal is input into the traditional shallow acoustic model, it usually undergoes only a few levels of linear or nonlinear transformation. Such shallow models currently have relatively mature structures and optimization algorithms. For more complex signals in speech tasks, by introducing deep models with stronger representation capabilities, the structure of the deep models is more complex and requires more levels of nonlinear transformations and has more advantages in expressing the nonlinear dynamic process of human speech. Strong modeling ability is also more suitable for processing complex time-varying signals; so, this article introduces a neural network structure LSTM with time "memory" function. LSTM controls the transmission of information in the network through a gating mechanism, which makes up for the defects of gradient disappearance and gradient explosion in the actual application of RNN, and has better performance in long-sequence speech tasks.

The gating mechanism of LSTM consists of forget gate, input gate, output gate, and memory unit. $f, i, o, C$ represent the state of forget gate, input gate, output gate, and cell memory unit, respectively. The forget gate can determine which historical information is forgotten or kept, and the output of the forget gate is expressed as

$$f_t = \sigma\left(W_f x_t + U_f s_{t-1} + b_f\right) f_t = \sigma\left(W_f x_t + U_f s_{t-1} + b_f\right), \quad (3)$$

where $W$ is the weight parameter from the input layer to the hidden layer, $U$ is the weight parameter from the hidden layer to the hidden layer, $b$ is the bias parameter, and $\sigma$ is the sigmoid function. At time $t$, the information of the hidden state at the previous time and the current input information are simultaneously input into the sigmoid function, and the output value is between 0 and 1. The closer to 0 means the more it should be forgotten, and the closer to 1 means the more it should be reserved.

The input gate is used to update the state of the memory unit. One part is the sigmoid layer, and the output value from 0 to 1 is used to determine which information to update. The equation is

$$i_t = \sigma(W_i x_t + U_i s_{t-1} + b_i). \quad (4)$$

The other part is the tanh layer, which uses output values from -1 to 1 to generate candidate update content. The equation is as follows:

$$C_t = \tanh\left(W_c x_t + U_c s_{t-1} + b_c\right). \quad (5)$$

Then, multiply the tanh output and the sigmoid output. The sigmoid output will determine which historical information in the tanh output is important and preserve it.

The output gate is used to determine which part of the information of the memory cell state is to be output. The output gate can determine the value of the next hidden state, which contains the relevant information of the previous input. The equation is as (6).

$$o_t = \sigma(W_o x_t + U_o s_{t-1} + b_o). \quad (6)$$

Before passing the output gate, the memory unit must pass the newly obtained state to the tanh function for processing and pass the new implicit state to the next time step. The equation is as follows:

$$s_t = o_t * \tanh\left(C_t\right). \quad (7)$$

## 5. Feature Selection and Parameter Extraction Based on Human Vocal Mechanism

Before extracting vocal timbre features, it was necessary to preprocess the vocal samples in the database in order to reduce the effects of aliasing and higher harmonic distortion caused by the human vocal organ and the equipment that collected the vocal signal.

*5.1. Preprocessing of Speech Signal.* The preprocessing process mainly includes preemphasis, framing, windowing, and endpoint detection.

*5.1.1. Preemphasis.* In the process of vocalization, there are many reasons for the weakening of the signal. Affected by glottal excitation and lip radiation, the high frequency part of the vocal signal frequency above 800 Hz will be greatly attenuated by 6 dB/octave, and the higher the frequency, the attenuation will be the more serious it is. In order to weaken this phenomenon, it is necessary to carry out preemphasis processing on the speech signal, so that the transition of the signal from the low frequency part to the high frequency part is more stable. This article adopts finite impulse

response (FIR) high-pass filter to realize its transfer function is defined as

$$H(z) = 1 - \mu z^{-1}, \tag{8}$$

where $\mu$ is the preemphasis coefficient, and the value range is $0.9 < \mu < 1.0$; in this experiment, $\mu = 0.96$.

*5.1.2. Framing and Windowing.* The speech signal is a time-varying signal, but due to the inertial movement of the vocal organs, within a very short time of 10-30 ms, the spectral and physical characteristics of the speech signal can be approximately considered to be stable. In this article, the method of using frame division and windowing was used in order to obtain a short-term stable speech signal, by using a window function to smoothly slide the speech signal, and cutting the speech signal by frame. The window length is the frame length. In order to ensure the continuity between the signal frame and the frame, the continuous overlapping frame method was used, and the overlap between the two frames before and after is the frame shift, which is generally half of the window length. The frame length set in this experiment is $M = 256$, and the frame shift is 128. The window function is defined as

$$\omega(n) = \begin{cases} 0.54 - 0.46 \cos\left[\dfrac{2\pi n}{(M-1)}\right], & 0 \le n \le M-1 \\ 0, & \text{other} \end{cases}. \tag{9}$$

*5.1.3. Endpoint Detection.* During the recording of voice samples, few pauses and adjustments are introduced when reading news articles. Therefore, there are some silent segments and background noise segments in the vocal data. This paper used a dual-threshold endpoint detection algorithm to identify and eliminate. The principle is as follows.

Assuming that the $n$-th frame of speech signal is $x_n(m)$, its short-term average energy is as equation (10).

$$E_n = \sum_{m=0}^{M-1} \left| x_n^2(m) \right|. \tag{10}$$

Short-term zero crossing rate is

$$z_n = \frac{1}{2} \sum_{m=0}^{M-1} \left| \operatorname{sgn}\left[x_n(m+1) - \operatorname{sgn}\left[x_n(m)\right]\right]\right|, \tag{11}$$

where $M$ is the frame length, and $\operatorname{sgn}[x]$ is the sign function.

When the signal-to-noise ratio of the vocal sample is large, the silent and voiced segments of the voice data can be distinguished by calculating the short-term average energy. When the signal-to-noise ratio does not meet the detection requirements, it is necessary to further calculate the average zero-crossing rate. When detecting a silent segment, a threshold is set. When the short-term energy of a frame is less than this threshold, it is judged as a silent frame; otherwise, it is a sound frame.

### *5.2. Parameter Extraction Based on Human Vocal Mechanism*

*5.2.1. Source Parameters.* This paper selected the pitch frequency that can reflect the vibration effect of the vocal cord as the source parameter. We used the short-term average amplitude difference function.

$$F_n(k) = \frac{1}{M} \sum_{n=0}^{M-k-1} x_i(M+k) - x_i(k), \tag{12}$$

where $x_i(n)$ is the voice signal of the $i$-th frame, and $k$ is the time delay.

*5.2.2. Vocal Tract Parameters.* In this paper, we chose the formant frequency and bandwidth as the vocal tract parameters, which can reflect the resonance quality and vocal tract model. Linear prediction cepstrum coefficient (LPCC) is also as the vocal tract parameters.

*(1) The Formant Frequency and Bandwidth.* The formant refers to the resonance frequency of the acoustic cavity. The number of formants reflects the fullness of the voice, and the position and bandwidth of the formants reflect the brightness and clarity of the voice. In this paper, the cepstrum method was used to extract the frequency and bandwidth of the first, second, and third formants. The specific steps are as follows:

(i) Perform Fourier transform on the $i$-th frame of speech signal $x_i(n)$. The equation is as follows:

$$X_i(k) = \sum_{n=0}^{N-1} x_i(n) e^{-j2\pi kn/N}. \tag{13}$$

(ii) Find the cepstrum of $X_i(k)$

$$\hat{x}_i(n) = \frac{1}{N} \sum_{n=0}^{N-1} \log_{10} X_i(k) e^{j2\pi kn/N}. \tag{14}$$

(iii) Cepstrum signal windowing: the equation is as follows:

$$h_i(n) = \hat{x}_i(n) \times h(n). \tag{15}$$

(iv) Find the envelope of $h_i(n)$

$$H_i(k) = \sum_{n=0}^{N-1} h_i(n) e^{-j2\pi kn/N}. \quad (16)$$

(v) Find the maximum value on the envelope of the speech spectrum and get the corresponding formant frequency and bandwidth

*(2) LPCC.* The generation of the speech signal can be regarded as the result of the convolution of the excitation signal and the frequency response of the vocal tract. To obtain the LPCC, it needs deconvolve by the method of homomorphic analysis and perform logarithmic processing. The specific steps for extracting LPCC are as follows:

(i) Perform Discrete Fourier Transform on the preprocessed speech signal $x(n)$. The equation is as follows:

$$X(k) = \sum_{-\infty}^{+\infty} x(n) \cdot \exp\left(-j \cdot \frac{2\pi}{N} \cdot n \cdot k\right). \quad (17)$$

(ii) Take the modulus and logarithm of the signal after discrete Fourier transform

$$C(k) = \ln\left[|X(k)|\right]. \quad (18)$$

(iii) Perform discrete cosine transform (DCT):

$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} C(k) \cdot \exp\left(j \cdot \frac{2\pi}{N} \cdot k \cdot n\right). \quad (19)$$

The calculation process of LPCC is small. It can not only describe the vowels well but also reflect the personality characteristics of the speech. However, LPCC is linearly approximate to the speech signal, which is contrary to the characteristics of human hearing and affects the performance of the system. So, it needs human hearing parameters as a supplement.

*5.2.3. Human Hearing Parameters.* To establish a high-performance objective evaluation system of timbre, in addition to study the sounding mechanism related to the timbre of broadcast speech, it is more important to simulate the mathematical model of timbre generation and the complex process of timbre perception based on the auditory characteristics of the human ear. Human hearing parameters mainly consider the subjective perception of human hearing and fully consider the characteristics of human hearing. They have a nonlinear logarithmic relationship with the frequency of the real speech signal, which is used to better characterize the nonlinear perception of human ears at different frequencies. Therefore, this paper selected cochlear frequency cepstrum coefficient (CFCC) and MFCC as the human hearing parameters.

*(1) CFCC.* The cochlea is an auditory receptor with a complex internal structure. The sensory hair cells can turn sound stimuli into nerve impulses, which are sensed through the auditory center of the brain. In addition, on the basement membrane of the cochlea, where sound peaks of different frequencies appear different, when the frequency is low, the peak value of the basement membrane appears near the cochlear foramen. As the sound frequency increases, the peak value of the basement membrane begins to move to the root, and the frequency of different sounds is logarithmically distributed on the basement membrane. The specific steps of CFCC extraction are as follows:

(i) Cochlear filter: the process of vocal signals passing through the cochlear filter simulates the frequency response of signals of different frequencies on the basilar membrane. The transformation process is as follows:

$$T(a, b) = \int_{-\infty}^{+\infty} f(t) \psi_{a,b}(t) dt, \quad (20)$$

where $f(t)$ is the speech signal function, and $\psi_{a,b}(t)$ is the cochlear filter function.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \left(\frac{t-b}{a}\right)^{\alpha} \exp\left(-2\pi f_L \beta\left(\frac{t-b}{a}\right)\right) \bullet \cos\left(2\pi f_L \left(\frac{t-b}{a}\right) + \theta\right) u(t). \quad (21)$$

$\alpha, \beta$, respectively, refers to the frequency domain shape and width of the cochlear filter function $\alpha = 3, \beta = 0.2$. u(t) is the unit step function. $\theta$ is the initial phase, $b$ is the Real numbers that can change over time, and $a$ is the scale variable and determined by the center frequency $f_c f_c$ and the lowest frequency $f_L$ of the cochlear filter bank.

(ii) The transmission mechanism of hair cells is defined as

$$h(a, b) = [T(a, b)]^2; \forall a, b. \quad (22)$$

The hair cell function of the $i$-th frequency band can be expressed as follows:

$$S(i, j) = \frac{1}{d} \sum_{b=l}^{l+d-1} h(i, b), l = 1, L, 2L, \cdots; \forall i, j, \quad (23)$$

where $L$ is the frame shift, and $d$ is the smoothing window length of the $i$-th frequency band.

(iii) Convert the output energy value of the hair cell function into perceived loudness through nonlinear cube root compression

$$y(i,j) = S(i,j)^{1/3}. \quad (24)$$

(iv) In order to remove the correlation between features, CFCC parameters are obtained after discrete cosine transform

*(2) MFCC*. MFCC is proposed based on the characteristics of human hearing. It is based on the linear cosine transform of the logarithmic power spectrum on the nonlinear Mel scale. The extraction steps of MFCC are as follows.

(i) Perform fast Fourier transform on the preprocessed speech signal $x(n)$ and obtain the discrete power spectrum $S(n)$ in the frequency domain. The obtained power spectrum is passed through a triangular Mel filter bank, so that the linear speech signal spectrum is converted into a Mel spectrum that can reflect the characteristics of human hearing. The output after passing $M = 126$ filter banks is $P_m$

(ii) Calculate the logarithmic energy of each filter bank output and take the logarithm of the output of the Mel filter bank. The equation is as follows:

$$I_m = \ln(P_m), 0 \le m \le M. \quad (25)$$

(iii) The discrete cosine transform is performed on the logarithmic energy obtained above, and the correlation between the characteristic parameters is reduced to obtain the 13th-order MFCC coefficient

(iv) In addition to extracting the static characteristics of MFCC coefficients, its dynamic characteristics are also analyzed. The extraction process is

$$D_n = \frac{\sum_{i=-k}^{k} i \cdot f(n+i)}{\sqrt{\sum_{i=-k}^{k} i^2}}. \quad (26)$$

where $f(n+i)$ represents the order of the $n+i$-th cepstral coefficient, and $k$ represents the interval of the difference frame generally taking 1 or 2.

*5.3. Feature Selection Strategy Based on Human Vocal Mechanism.* In the process of objective evaluation of broadcasting vocal timbre, the selection of feature parameters usually is multidimensional to ensure the evaluation. In addition, there are many organs of control and cooperation, such as respiratory organs, vocal organs, resonance organs, and nervous system [26]. Therefore, it is very important to extract the appropriate feature parameters from the complex and diverse feature parameters to characterize the timbre feature completely and accurately. Preliminary experiments show that it is not that the more dimensions of the feature parameters are selected, the higher accuracy of the resulting classification is got. On the contrary, more dimensions will increase unnecessary calculations. For the feature parameters, the redundancy and correlation should be removed as much as possible in the dimension reduction. And it is also necessary to characterize the timbre features in combination with the mechanism of human vocal production. Therefore, the optimized feature selection module is combined with human vocal generation mechanism and PCA. Figure 3 shows the feature selection process.

*5.3.1. Mechanism of Human Vocal.* Broadcasting and hosting vocal production is a kind of purposeful artistic vocal, which originates from the spoken vocal in daily life, but it is more professional than that. In the process of broadcasting, perfect cooperation is needed for respiratory organs, vocal organs, and resonance organs to achieve good effect of the emotional expression. In order to establish a high-performance objective evaluation system of timbre, it is necessary to study the mechanism of sound production related to timbre and the auditory feature of human ears.

*(1) Respiratory Organs*. Respiratory organs are composed of lung, chest, diaphragm, abdominal muscle, trachea, and bronchus, which are the "source" power in vocal production. Figure 4(a) shows the specific structure.

After inhalation, the breath is distributed to the alveoli of the left and right lobes through the trachea and bronchus and then enters the bronchus from the outlet of the lung and converges to the trachea. The vocal cords vibrate by mobilizing the muscles of the chest and abdomen. The breath exhaled from the lung is the driving force of human vocal. The strength of human vocal and the resonance state are directly related to the speed of exhaled air flow and the size of flow pressure. Therefore, the respiratory organ is the power and energy guarantee of broadcasting, and the movement of muscle groups related to respiration also plays an important role.

*(2) Vocal Organs*. The vocal organ is composed of the larynx and vocal cord. The larynx is a complex system of cartilage and muscle that controls the movements of vocal cord, including cricoid cartilage, thyroid cartilage, arytenoid cartilage, lower airway, and upper pharyngeal cavity. The vocal cord is in the middle of the larynx. It is two symmetrical and elastic white ligaments juxtaposed horizontally. It is regulated by the cartilage and muscle in the larynx. Figure 4(b) shows the specific structure.

The vibration process of the vocal cord is related not only to the air velocity and pressure of breathing but also to the control of cartilage and muscle groups of the larynx.

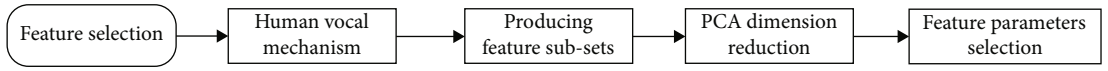


Figure 3: Feature selection process.

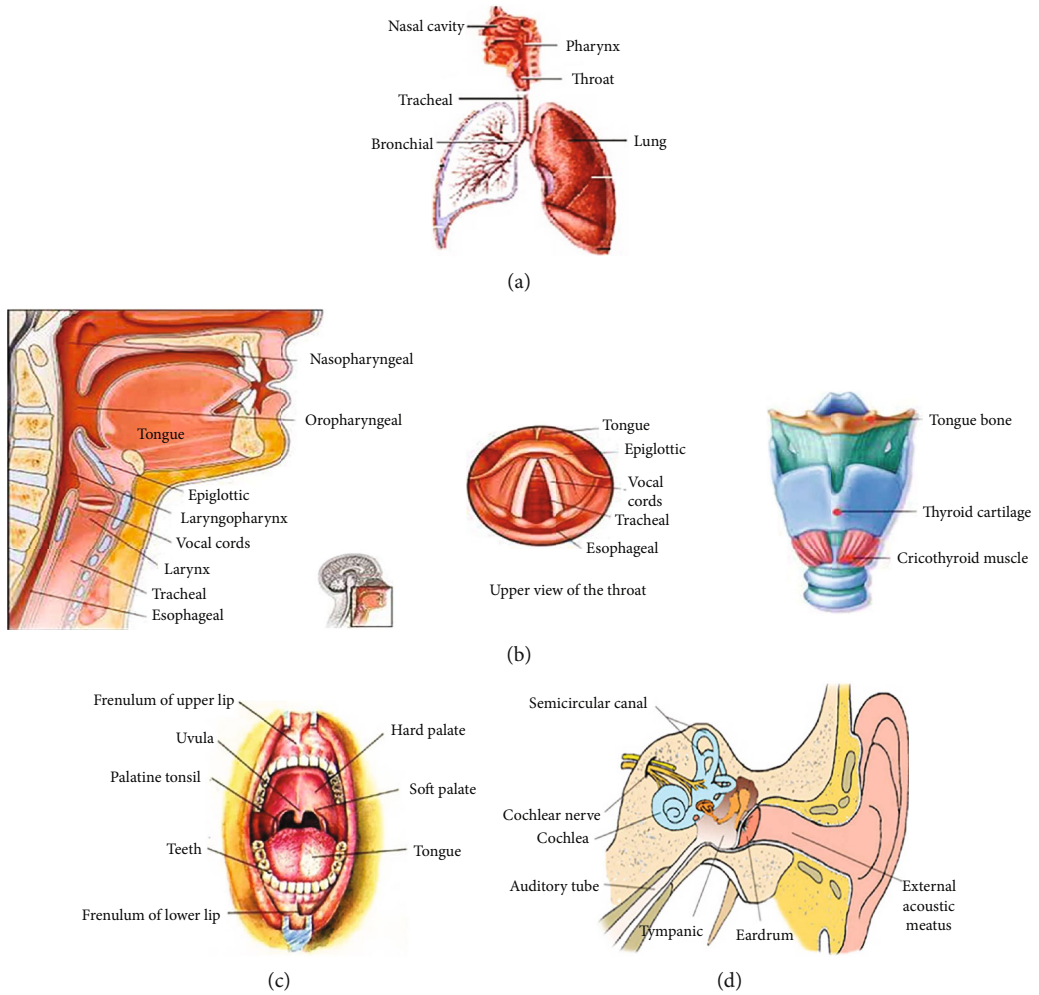


Figure 4: Composition of respiratory organs. (a) Structural diagram of respiratory organs. (b) Structural diagram of vocal organs. (c) Structure of resonance organ. (d) Structural diagram of auditory organs.

In the process of vocal broadcasting, it is necessary to fully mobilize the cartilage and muscle groups of the larynx, to control the breath and vocal organs to make the sound production stable and free.

*(3) Resonance Organ*. The resonance organ is composed of chest cavity, head cavity, and oral cavity. The chest cavity includes trachea, bronchus, and the whole lung. The head cavity includes nasal cavity, maxillary sinus, frontal sinus, and sphenoid sinus. The oral cavity includes upper and lower lips, upper and lower teeth, upper and lower gingiva, upper and lower jaws, tip of tongue, surface of tongue, root of tongue, and small tongue. The sound produced by vocal cords can obtain complex and colorful timbre only by adjusting the resonators. Oral cavity is the main speech resonator and manufacturer of various timbres. Figure 4(c) shows the specific structure.

*(4) Auditory Organ*. The auditory organ is composed of the external ear, the middle ear, and the inner ear. The outer ear includes the auricle, the external auditory canal, and the eardrum, and it is mainly used to collect sound, identify sound sources, and expand the sound of certain frequencies; the middle ear includes the malleus, incus, and stapes, which mainly transmits sound from the external auditory canal to the cochlea. The comprehensive function of the external ear and the middle ear is equivalent to a smooth band filter between 500 Hz and 6 KHz. The inner ear includes vestibular window, circular window, oval window, and cochlea. The main function of the cochlea is like a spectrum analyzer,

which can decompose complex signals into various frequency components. Figure 4(d) shows the specific structure.

According to the coordination of the related organs in vocal broadcasting, the selected features can be divided into three categories: source parameters reflecting the performance of respiratory organs, vocal tract parameters reflecting the performance of vocal organs and common vocal organs, and human hearing parameters reflecting the process of timbre perception. The parameters of feature classification are shown in Table 2.

*5.3.2. Basic Mechanism of PCA.* This paper mainly studies the individual characteristic parameters that can characterize the difference in timbre of the speaker during the process of broadcasting and vocalization. According to the principle of human vocalization and the auditory characteristics in the process of timbre perception, the commonly used feature parameters are divided into three categories, which include source parameters, vocal tract resonance parameters, and human hearing parameters. In the process of timbre classification, the input sample is a vector set of multiframe feature parameters of the entire speech. In the process of repeated training, there is a great correlation between frames. There is also redundant information between each feature parameter. In order to better retain the most effective and most reflective information of personalized timbre characteristics, PCA technology is utilized to perform feature optimization for selecting the effective components in the timbre characteristics. The basic mechanism of algorithm PCA is as follows:

(1) Calculate the mean value $\mu_i$ of the timbre feature vector

If each timbre feature sample dimension is set to, then each sample data can be expressed as

$$X = \left(x_1, x_2, x_3 \cdots, x_p\right)^T. \tag{27}$$

If the sample set has $N$ pieces of sample data, the mean value of the sample set can be expressed as

$$\mu_i = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{28}$$

(2) Calculate the covariance matrix $S_i$ of the timbre features as follows:

$$S_i = \frac{1}{N-1}\left(X_{\mathrm{i}} - \mu_i\right)^T \left(X_i - \mu_i\right). \tag{29}$$

(3) Calculate the eigenvalues $\lambda_1, \lambda_2, \lambda_3, \cdots \lambda_i$ and its corresponding eigenvector $w_1, w_2, w_3, \cdots w_i$ of the covariance matrix $S_i$ and sort them in descending order

(4) Reduce the timbre feature dimension to $k$, and a new feature matrix $K$ is formed by obtaining eigenvectors with larger eigenvalues. Matrix transformation is showed in Equation (30). The result $Y_i$ is the set of timbre features after dimension reduction

$$Y_i = K^T X_i. \tag{30}$$

The cumulative contribution rate is used to select the principal component in determining the feature dimension. The eigenvalues $\lambda_1, \lambda_2, \lambda_3, \cdots \lambda_i$ are sorted from large to small, obtained from the abovementioned feature parameters. The ratio of the sum of the first $k$ principal components to the total eigenvalue is the cumulative contribution rate of the first $k$ principal components. Finally, the feature parameters are selected with the cumulative contribution rate greater than 85%.

*5.4. Results of Timbre Feature Selection.* Feature selections are generated according to the mechanism of human vocal generation, and all features are classified into three types, which include source parameters, vocal tract parameters, and human hearing parameters. Then, the dimensions of various feature combinations are reduced by PCA technology. Finally, the selected source parameters are pitch frequency, vocal tract parameters are formant frequency, bandwidth, and LPCC, and human hearing parameters are CFCC and MFCC. Table 3 shows the feature selection results.

In order to further verify the scientificity of feature selection, we conducted further experiments with unfiltered original features as a control group and compared the differences between filtered and unfiltered timbre feature parameters in the timbre objective evaluation model. Figure 5 shows the results.

It can be seen from Figure 5 that regardless of the male or female samples, in the timbre objective evaluation models of HMM, GMM-UBM, and LSTM, the timbre characteristic parameters after screening are more accurate than the original characteristic parameters classification. This further verifies the scientificity and necessity of selecting acoustic parameters based on the principle of human vocalization in the objective evaluation system of timbre.

## 6. Objective Evaluation Method

Before the experiment based on objective evaluation method, details of the experiment data, evaluation index, and experimental design were as follows.

(1) Experiment data: the speech samples in the database were processed uniformly. The preemphasis coefficient was set to 0.96, the frame length was 256 sampling points, and the frame shift was 128 sampling points. Hamming window was used to extract the vocal feature parameters of each speech sample,

Table 2: Feature classification parameters.

| | |
|---|---|
| Source parameters | Short-term average energy, short-term average amplitude, pitch frequency, short-term energy change rate, long-term average energy, fundamental frequency change rate |
| Vocal tract parameters | Formant frequency, formant bandwidth, linear prediction cepstrum coefficient, line spectrum frequency |
| Human hearing parameters | Mel frequency cepstrum coefficient, cochlear frequency cepstrum coefficient, and perceptual linear prediction coefficient |

Table 3: Feature selection results.

| Source parameters | Pitch frequency |
|---|---|
| Vocal tract parameters | The frequency and bandwidth of the first, second, and third formants, linear predictive cepstrum coefficient |
| Human hearing parameters | Mel frequency cepstrum coefficient and cochlear frequency cepstrum coefficient |

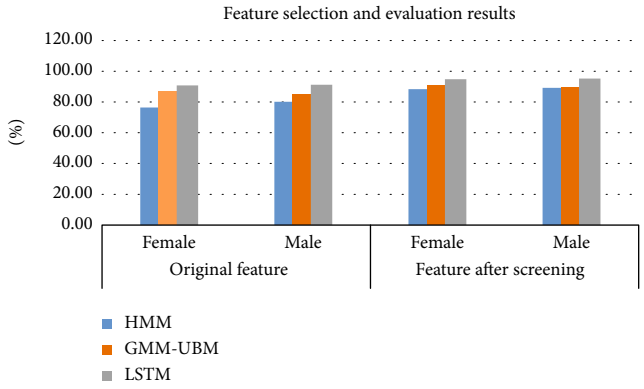


Figure 5: Feature selection and evaluation results.

and the 74-dimensional feature parameters were stored in an Excel table

(2) Evaluation index: 90% of the database was taken as the training selection, and the remaining 10% was taken as the test selection. The results were based on the speech samples of each 20 seconds test selection, and the accuracy of the broadcasting vocal timbre grading Acc is calculated by

$$\text{Acc} = \frac{Q}{U} \times 100\%, \tag{31}$$

where $Q$ represents the number of speech samples with correct thin thickness classification in the test selection, and $U$ represents the total number of speech samples contained in the test selection.

(3) Experimental design: the feature selection module based on the human vocal mechanism and PCA was used to divide the acoustic features into source parameters, vocal tract parameters, and human hearing parameters. The dimension of each feature type was reduced by PCA, and we conducted fusion and nonfusion experiments on the three types of optimal features, respectively. Among them, we used four timbre feature combinations for fusion. The inputs were 74-dimensional feature parameters, and the thin thickness of broadcasting vocal timbre was divided into three grades, which are *A*, *B*, and *C*. *A* is excellent, *B* is good, and *C* is pass. The accuracy of the final classification results was used to evaluate the selected three objective evaluation methods; so, this can explore a better objective evaluation method of broadcasting vocal timbre. Table 4 shows the relevant information of the experiment

### 6.1. Timbre Objective Evaluation Method Based on HMM

*6.1.1. Experiment Workflow.* Figure 6 shows the experiment workflow of the objective evaluation model based on HMM, and it mainly includes speech signal preprocessing, vocal timbre feature parameters extraction, HMM training parameters, and Viterbi algorithm recognition.

*6.1.2. Analysis of Experimental Results.* The source parameters, vocal tract parameters, and human hearing parameters were, respectively, analyzed on three types of characteristic parameters and their different characteristic combination types for objective evaluation of timbre. The accuracy of timbre classification was obtained by implementing the objective evaluation model based on HMM. Figure 7 shows the results.

The radar chart in Figure 7 shows that HMM obtains good accuracy in the objective evaluation of broadcast timbre. From the perspective of feature fusion, the grading effect obtained by using a single timbre characteristic parameter is not as good as the fusion timbre characteristic parameters. Among them, only the source parameters are used as the input, and the accuracy rate is only 77.25%. The classification effect obtained by the fusion of the three types of features as input is the best, and the accuracy rate can reach up to 89.28%, which further verifies the scientific nature of using the human vocal mechanism as the basis for the selection of timbre features. From the perspective of gender

Table 4: Experiments related information.

| Speech language type | The number of samples | The number of training samples | The number of testing | Sampling frequency | Window type | Window size | Feature dimension | Fusion type | Classifier type |
|---|---|---|---|---|---|---|---|---|---|
| Chinese | 6000 | 5400 | 600 | 44.1 KHz | Hamming window | 256 | 74-dimensional | Early fusion | Discrete |

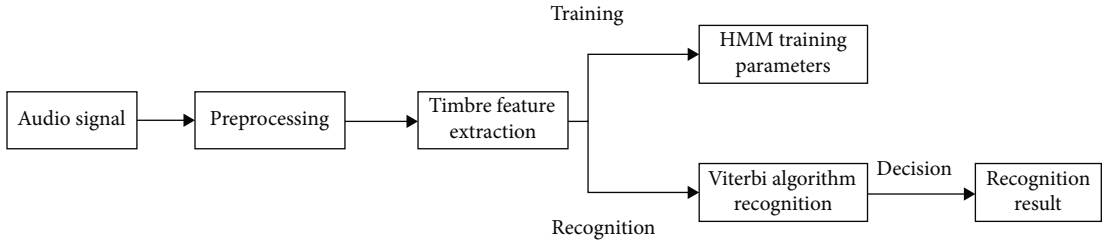


Figure 6: Workflow of the objective evaluation model based on HMM.

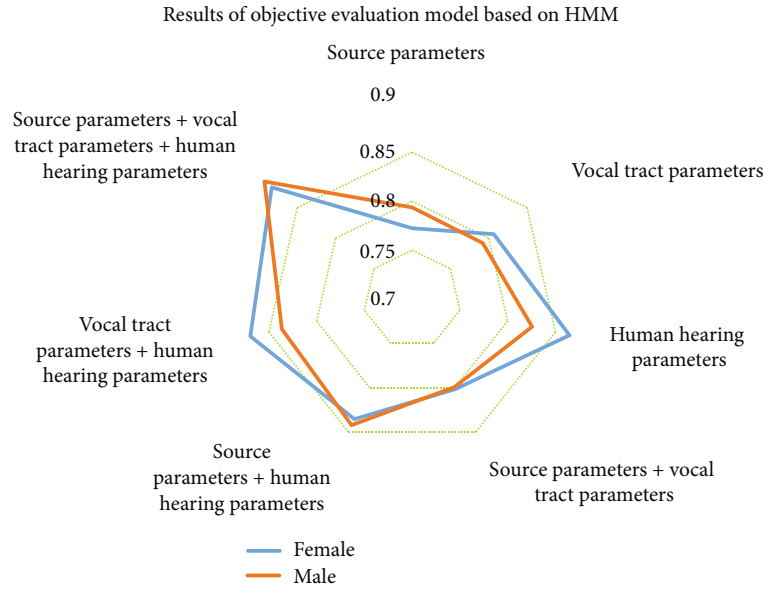


Figure 7: Results of the objective evaluation model based on HMM.

differences in timbre, when the same feature parameters combination is used as input, girls' timbre classification accuracy rate is generally higher than that of men. Especially when human hearing parameters are used as input, the accuracy rate of girls is about 4% higher than that of boys. Maybe it is because girls' timbre is more recognizable in terms of auditory perception.

### 6.2. Timbre Objective Evaluation Based on GMM-UBM

*6.2.1. Experiment Workflow.* Figure 8 shows the experiment workflow of the objective evaluation model based on the Gaussian Mixture Model-Universal Background Model (GMM-UBM), and it mainly has three important models, UBM training parameters, MAP self-adaption, and recognition result.

*6.2.2. Network Parameter Setting*

(i) The order of GMM model: in this experiment, 100 speech samples are selected, and 74-dimensional vocal timbre feature parameters are extracted, respectively. The training speech sample duration is 20 seconds, the testing speech sample duration is 6 seconds, and the GMM order is 4, 8, 16, 32, and 64. Table 5 shows the results

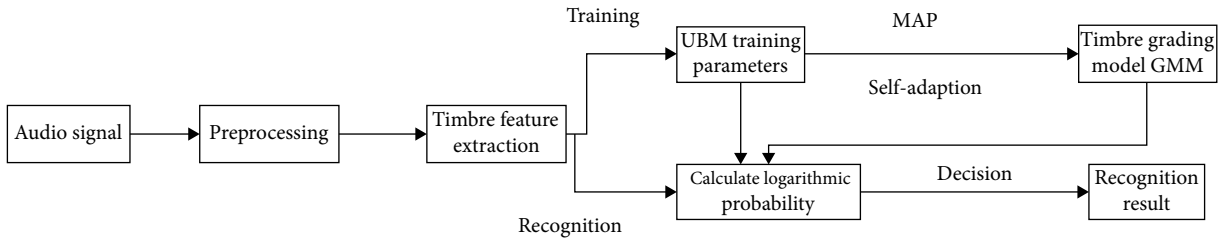


Figure 8: Workflow of the objective evaluation model based on GMM-UBM.

Table 5: Results of GMM model order.

| GMM order | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|
| Classification accuracy | 60.21% | 71.35% | 85.63% | 89.48% | 89.98% |
| Training duration | 568.7 s | 663.9 s | 1035.1 s | 1335.6 s | 2147.3 s |

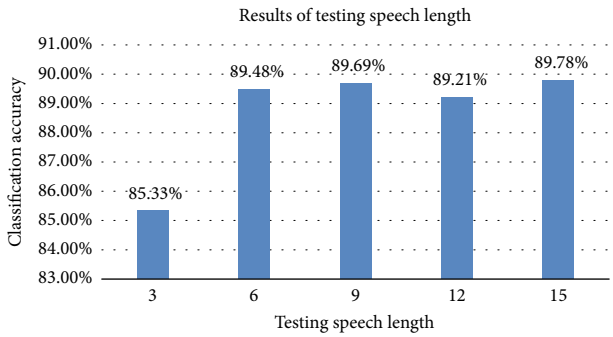


Figure 9: Results of testing speech length.

(ii) Length of the testing speech: in this experiment, 100 speech samples are selected, the GMM order is 32, the training speech sample duration is still taken as 20 seconds, and the testing speech sample is intercepted for 3 s, 6 s, 9 s, 12 s, and 15 s for comparative experiments. Figure 9 shows the results

The results show that with the increase of GMM order, the accuracy of 64 order classification is not obviously improved compared with that of 32 order classification; however, the training time is close to 2 times of that of 32 orders. In order to balance the relationship between classification accuracy and computational complexity, the order of GMM is selected as 32. When test speech length is greater than 6 s, the improvement of accuracy becomes very insignificant with the increase of test speech length. Because the test speech is too long, the model may not be able to cover and match the recognition process. Therefore, the test speech length was chosen as 6 s.

*6.2.3. Analysis of the Results.* The source parameters, vocal tract parameters, and human hearing parameters were, respectively, analyzed on three types of characteristic parameters and their different characteristic combination types for objective evaluation of timbre. The accuracy of timbre classification was obtained by inputting objective evaluation model based on GMM-UBM. Figure 10 shows the results.

The results show that the improved GMM-UBM generally improves the vocal timbre classification effect compared with the traditional HMM, but it still has the feature fusion problem as HMM. Overall, the fusion feature is better than the unfused feature classification effect. When source parameters, vocal tract parameters, and human hearing parameters are used as input, the accuracy of the objective evaluation of women's timbre can reach 90.5%. In addition, when human hearing parameters are used as input, the accuracy of objective evaluation of both male and female timbre reaches more than 90%. This shows the scientific nature of selecting timbre parameters based on the human vocal mechanism.

### 6.3. Timbre Objective Evaluation Based on LSTM

*6.3.1. Experiment Workflow.* The key point of the objective evaluation based on long short-term memory (LSTM) is to estimate the weight parameters of gate structure in the network training. Figure 11 shows the experiment workflow of the objective evaluation model based on LSTM.

*6.3.2. Network Parameter Setting.* The size of the hidden layer and the number of iterations has different influence degrees on the model classification accuracy in the training process. Therefore, this paper makes a comparative analysis of the size of the hidden layer and the number of iterations of the LSTM network through the design of preliminary experiments, to determine several important parameters affecting the network performance.

(i) Number of iterations: using the LSTM network model, the initial learning rate is 0.05, the learning rate decline cycle is 50, the decline factor is 0.2, and the dropout parameter is 0.2. The timbre feature data

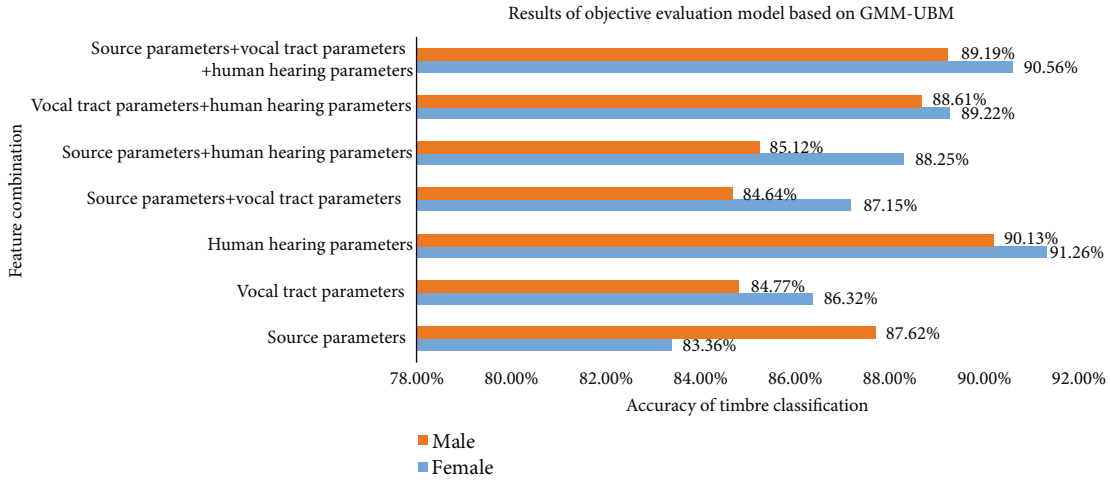


Figure 10: Results of objective evaluation model based on GMM-UBM.

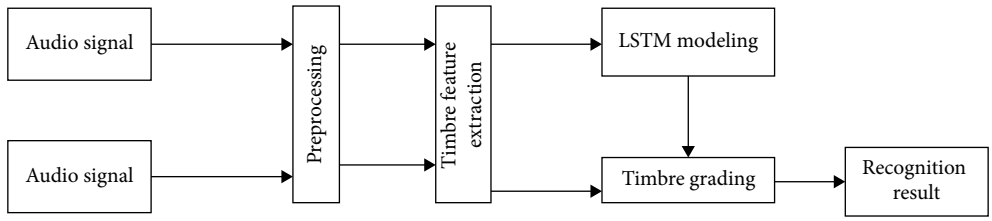


Figure 11: Workflow of objective evaluation model based on LSTM.

Table 6: Results of iteration times.

| Number of iterations | 25 | 50 | 75 | 100 | 125 | 150 | 175 | 200 |
|---|---|---|---|---|---|---|---|---|
| Accuracy/% | 89.21 | 91.54 | 92.83 | 93.66 | 94.21 | 94.89 | 94.97 | 95.01 |
| Training duration/s | 1903.8 | 2509.4 | 3147.9 | 3706.9 | 4387.2 | 5006.7 | 5701.9 | 6421.8 |

are input into the model, and the iteration times are set as 25, 50, 75, 100, 125, 150, 175,and 200. The evaluation indexes such as classification accuracy and training time are obtained by model training, and the appropriate iteration times are obtained by comparison. Table 6 shows the results

(ii) Hidden layer scale: using the LSTM network model, the initial learning rate is 0.05, the learning rate decline cycle is 50, the decline factor is 0.2, the dropout parameter is 0.2, and the iteration times are set to 150. The timbre feature data are input into the model, and the hidden layer nodes are set as 50, 100, 150, 200, 250, and 300. The evaluation indexes such as classification accuracy and training time are obtained by comparison. Table 7 shows the results

The results show that when the number of iterations is more than 150, the accuracy is no longer significantly improved, but the training time is still greatly increased. Considering the accuracy of classification, the system operation time is shortened as much as possible; so, the iteration number of LSTM network is selected as 150.When the number of hidden layer nodes continues to 300, the accuracy rate begins to decline to a certain extent. Because the number of neurons is more, the convergence speed of the model is slower. Considering the premise of meeting the classification accuracy, the number of hidden layer nodes is 200 in LSTM network.

*6.3.3. Analysis of Experimental Results.* The source parameters, vocal tract parameters, and human hearing parameters were, respectively, analyzed on three types of characteristic parameters and their different characteristic combination types for objective evaluation of timbre. The accuracy of timbre classification was obtained by inputting the objective evaluation model based on LSTM. Figure 12 shows the results.

The results show that the LSTM network using the deep learning algorithm has better classification effect than the

Table 7: Results of hidden layer node number.

| Number of hidden layer nodes | 50 | 100 | 150 | 200 | 250 | 300 |
|---|---|---|---|---|---|---|
| Accuracy/% | 93.64 | 93.98 | 94.21 | 94.97 | 94.32 | 93.57 |
| Training duration/s | 2631.7 | 3409.4 | 4194.2 | 5094.8 | 5901.6 | 8004.9 |

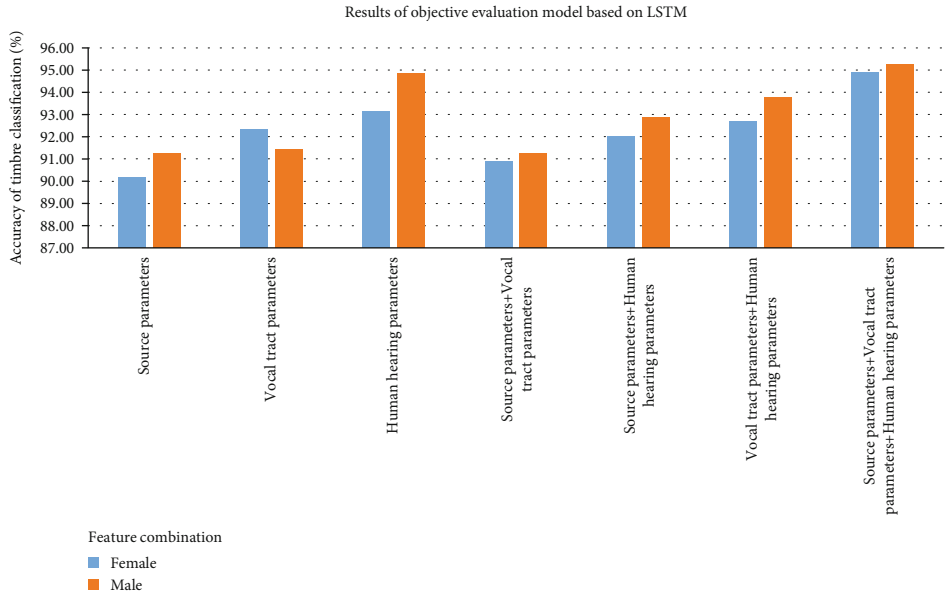


Figure 12: Results of the objective evaluation model based on LSTM.

traditional HMM and the improved GMM-UBM in the task of vocal timbre classification. Regardless of whether it is a fusion of timbre characteristic parameters or not, the accuracy of objective evaluation of LSTM-based broadcast timbre is above 90%. When the source parameters, vocal tract parameters, and human hearing parameters are used as input, the overall accuracy of objective evaluation of timbre is about 5% higher than that of HMM and GMM-UBM. This is because the LSTM network continues the advantages of the recurrent neural network in dealing with timing problems, and it still has a good expression ability for the long-term speech classification.

## 7. Conclusions

This paper is based on the background of the selection and training of broadcasting talents by various professional colleges and recruitment units. Linguistics, journalism, and psychology form the basis of our research. A series of studies have been carried out to get a set of objective grading methods for vocal timbre with scientific evaluation and flexible operation. Combining the mechanism of human vocalization and PCA technology, we divide the characteristic parameters of broadcast timbre into three categories, which include source parameters, vocal tract parameters, and human hearing parameters. The objective evaluation experiment of broadcast timbre is carried out by using the combination of these timbre characteristics. Experiments show that the selection of timbre parameters based on the human vocal mechanism is scientific. Furthermore, the objective evaluation method based on LSTM has better classification effect than the traditional HMM and GMM-UBM in the timbre classification. The LSTM method introduces a deep model for complex audio signals with stronger representation ability and more complex structure. It can better simulate the nonlinear dynamic process of human vocal with levels of nonlinear transformation, and it has excellent performance in the vocal timbre classification. The classification accuracy rate is about 95%.

In the future, we intend to improve the performance of the objective evaluation system of timbre through different fusion technologies, such as late fusion based on score fusion, namely, mean, maximum, and linear weighted sum fusion [24], and plan to combine the *i*-vector approach to improve the accuracy of the timbre objective evaluation system [23]. In addition, we are currently collecting a large timbre speech evaluation database containing other categories to further improve the proposed model.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the National Key R&D Program of China (2018YFB1404100).

## References

[1] W. He, Y. Zou, and Y. H. Ma, "CUCBNC: a broadcast news voice database introducing broadcasting knowledge," *Journal of Tsinghua University*, vol. 51, no. 9, pp. 1313–1316, 2011.

[2] S. Rawat, T. Khan, and A. Chauhan, "PRAAT software; utilization of computerized approach for determination of variation present in recorded audios from distinct sources," *International journal of engineering trends and technology*, vol. 67, no. 3, pp. 111–114, 2019.

[3] K. L. Spreadborough and I. Anton-Mendez, "It's not what you sing, it's how you sing it: How the emotional valence of vocal timbre influences listeners' emotional perception of words," *Psychology of Music*, vol. 47, no. 3, pp. 407–419, 2019.

[4] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, article 107020, 2020.

[5] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[6] A. Pandey and D. L. Wang, "On cross-corpus generalization of deep learning based speech enhancement," *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 28, pp. 2489–2499, 2020.

[7] F. Pishdadian, G. Wichern, and J. Le Roux, "Finding strength in weakness: learning to separate sounds with weak supervision," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2386–2399, 2020.

[8] G. Gosztolya, "Posterior-thresholding feature extraction for paralinguistic speech classification," *Knowledge-Based Systems*, vol. 186, article 104943, 2019.

[9] T. Zoughi, M. M. Homayounpour, and M. Deypir, "Adaptive windows multiple deep residual networks for speech recognition," *Expert Systems with Applications*, vol. 139, article 112840, 2020.

[10] X. Zhang, X. Zou, M. Sun, T. F. Zheng, C. Jia, and Y. Wang, "Noise robust speaker recognition based on adaptive frame weighting in GMM for I-vector extraction," *IEEE Access*, vol. 7, pp. 27874–27882, 2019.

[11] M. T. S. Al-Kaltakchi, R. R. O. Al-Nima, and M. A. M. Abdullah, "Comparisons of extreme learning machine and backpropagation-based i-vector approach for speaker identification," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 3, pp. 1236–1245, 2020.

[12] S. Yang, S. Arif, M. Jia, and S. Zhong, "SAL-Net: Self-supervised attribute learning for object recognition and segmentation," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 2891303, 13 pages, 2021.

[13] M. Yan, X. Lou, and Y. Wang, "Channel noise optimization of polar codes decoding based on a convolutional neural network," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 1434347, 10 pages, 2021.

[14] A. Aggarwal, A. Srivastava, A. Agarwal et al., "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, 2022.

[15] M. B. Akçay and K. Oğuz, "Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[16] M. Yan, S. Li, C. A. Chan, Y. Shen, and Y. Yu, "Mobility prediction using a weighted Markov model based on mobile user classification," *Sensors*, vol. 21, no. 5, 2021.

[17] M. Yokoyama, "Possibility of distinction of violin timbre by spectral envelope," *Applied Acoustics*, vol. 157, article 107006, 2020.

[18] C. Jin, Z. Song, J. Xu, and H. Gao, "Attention-based Bi-DLSTM for sentiment analysis of Beijing opera lyrics," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 1167462, 8 pages, 2022.

[19] R. Pradeep and K. S. Rao, "Incorporation of manner of articulation constraint in LSTM for speech recognition," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3482–3500, 2019.

[20] Y. Zhuo, L. Yan, W. Zheng, Y. Zhang, and C. Gou, "A novel vehicle detection framework based on parallel vision," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 9667506, 11 pages, 2022.

[21] J. U. Bang, M. Y. Choi, S. H. Kim, and O. Kwon, "Automatic construction of a large-scale speech recognition database using multi-genre broadcast data with inaccurate subtitle timestamps," *IEICE Transactions on Information and Systems*, vol. -E103.D, no. 2, pp. 406–415, 2020.

[22] M. Al-Qaderi, E. Lahamer, and A. Rad, "A two-level speaker identification system via fusion of heterogeneous classifiers and complementary feature cooperation," *Sensors*, vol. 21, no. 15, p. 5097, 2021.

[23] M. T. S. Al-Kaltakchi, M. A. M. Abdullah, W. L. Woo, and S. S. Dlay, "Combined i-vector and extreme learning machine approach for robust speaker identification and evaluation with SITW 2016, NIST 2008, TIMIT databases," *Circuits, Systems, and Signal Processing*, vol. 40, no. 10, pp. 4903–4923, 2021.

[24] M. T. S. Al-Kaltakchi, W. L. Woo, S. Dlay, and J. A. Chambers, "Evaluation of a speaker identification system with and without fusion using three databases in the presence of noise and handset effects," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, Article ID 80, pp. 1–17, 2017.

[25] R. Jahangir, Y. W. Teh, N. A. Memon et al., "Text-Independent speaker identification through feature fusion and deep neural network," *IEEE Access*, vol. 8, pp. 32187–32202, 2020.

[26] M. Yan, W. Li, C. A. Chan, S. Bian, I. Chih-Lin, and A. F. Gygax, "PECS: towards personalized edge caching for future service-centric networks," *China Communications*, vol. 16, no. 8, pp. 93–106, 2019.