Entertainment Computing 52 (2025) 100814

Contents lists available at ScienceDirect

Entertainment Computing

journal homepage: www.elsevier.com/locate/entcom

# Social robot assisted music course based on speech sensing and deep learning algorithms

Xiao Dan

*College Of Music, Qilu Normal University, Jinan 210031, China*

ARTICLE INFO

*Keywords:*
Speech sensing
Deep learning algorithms
Social robots
Course in music

ABSTRACT

In the field of social robot teaching, research has focused on how to use technological means to provide better learning support and personalized interactive experiences. Social robots can interact with students and provide personalized learning support, thereby improving their learning effectiveness and engagement. The speech sensing model of social robots can perceive students' emotions and feedback in real-time through technologies such as speech recognition and sentiment analysis, thereby providing intelligent responses and guidance. The deep learning recommendation model for music course resources extracts music features through deep learning techniques, and combines session interest extraction techniques to personalized recommend music resources suitable for students' interests and abilities. By analyzing students' interests and learning goals, robots can provide music learning resources that meet their needs based on recommendation algorithms, further stimulating their learning interest and enthusiasm. The experimental results show that the use of social robots in the learning environment significantly improves the learning effectiveness and participation of students. Through personalized interaction and intelligent response guidance, students are more likely to understand and master music knowledge, while experiencing joyful and positive learning emotions. The study validated the effectiveness of social robot assisted music courses based on speech sensing and deep learning algorithms, demonstrating its advantages in improving student learning effectiveness and engagement.

## 1. Introduction

In the field of music education, the application of social robots in assisting music course teaching has become a new teaching model, bringing new possibilities to music education [1]. In traditional music classrooms, students usually need to follow the teacher's guidance for practice and learning. But with the emergence of social robots, music teaching can achieve a new way of interaction [2]. Social robots utilize voice sensing technology to perceive students' voice commands and respond accordingly, increasing their participation in music education [3]. Deep learning algorithms can help social robots better understand the needs and feedback information of students [4]. By analyzing and learning from a large amount of data, social robots can provide more personalized and effective teaching services based on the characteristics and learning progress of individual students [5].

The application of social robots in music course teaching is an innovative teaching model that leverages the advantages of voice sensing and deep learning algorithms, bringing new possibilities to music education [6]. With the application of voice sensing and deep learning algorithms, the social robots designed in this article have become effective auxiliary tools in music education [7]. They can provide personalized guidance and interactive experiences, helping students better understand and master music knowledge and skills. Traditional music courses often follow a unified schedule and teaching content, while social robots can provide personalized teaching based on students' rhythm and interests. Students can learn at their own pace and are no longer limited by traditional classroom modes. Traditional music education typically requires a significant amount of manpower and material investment, including teacher training and curriculum development. Social robots can to some extent alleviate the burden on teachers, improve teaching efficiency, and give teachers more time and energy to focus on personalized tutoring and provide professional guidance.

## 2. Related work

The literature has designed a creative evaluation index system for robot education scenarios [8]. The study initially designed evaluation

*E-mail address:* 18615419620@163.com.

https://doi.org/10.1016/j.entcom.2024.100814
Received 22 May 2024; Received in revised form 26 June 2024; Accepted 8 July 2024
Available online 9 July 2024

indicators, and then used the Delphi method for iteration, improvement, and optimization of evaluation indicators. Finally, a complete evaluation indicator system was determined to construct a creativity perspective measurement table for educational robot activities. The creativity evaluation index is a tool used to measure the level of creativity exhibited by students in educational robot activities [9]. Creativity refers to the individual's ability to generate unique and innovative solutions when facing problems or challenges. When designing the evaluation index system, the literature used the Delphi method. The Delphi method is a method of expert discussion and consensus, which gradually improves and optimizes evaluation indicators through multiple rounds of investigation and feedback [10]. A group of experts, including education experts, psychologists, and experts in the field of robotics education, were invited to discuss and comprehensively evaluate evaluation indicators. In each round of Delphi survey, experts rate and provide feedback on evaluation indicators. Literature revises and optimizes indicators based on expert feedback, and invites experts again for evaluation, iterating continuously until a consensus is reached. The final evaluation index system includes multiple aspects, such as thinking flexibility, problem-solving ability, innovative thinking, originality, etc. Each indicator has a clear definition and scoring criteria, which can be used to evaluate the creative performance of students in educational robot activities [11]. These evaluation indicators combine the needs of educational robot activities and the characteristics of student creativity, and have good guiding significance and practical value. The literature introduces a robot social robot intelligent voice interaction system, which has online and offline voice interaction functions [12]. By designing and implementing each module, and integrating them through ROS (Robot Operating System), a complete robot social robot intelligent voice interaction system is ultimately formed. The voice wake-up module in the literature is used to detect the user's voice input and trigger the system response [13]. It can wake up based on the characteristics of the voice and transfer control to the next module after recognizing the user's voice. The intention recognition module of literature is used to understand the user's intention and needs [14]. It analyzes and processes textual information, extracts the user's intention, and performs corresponding operations or provides corresponding answers as needed. The speech synthesis module of the literature is responsible for converting the system's answers into speech information [15]. Based on the output and feedback of the system, this module can convert text information into speech, enabling the system to have natural and smooth voice interaction with users. Through the integration and operation of ROS, the literature enables various modules to communicate and coordinate with each other, forming a complete robot social robot intelligent voice interaction system. This system can be applied to fields such as robot social interaction and intelligent assistants, providing users with convenient and intelligent voice interaction experiences.

The literature introduces a method for extracting feature parameters of intelligent robot speech signals based on VMD (Variational Mode Decomposition) [16]. VMD technology is used to establish pulse digital models of speech signals, analyze real-time system frequencies, and determine signal transfer functions. By treating intelligent robots as large biomolecules for feature parameter analysis of speech signals, the literature adopts the molecular programming mode unique to VMD and establishes a molecular visualization program to split, analyze, and process unstable speech signals [17]. The literature uses VMD technology to decompose speech signals into multiple modalities. This decomposition method can decompose complex speech signals into a series of modalities, each representing a specific frequency range of signal components. By analyzing each modality, speech features within a specific frequency range can be extracted. The literature models and analyzes the characteristics of speech signals in intelligent robot application scenarios. Due to the instability and variability of speech signals in intelligent robots, literature considers them as large biomolecules and establishes corresponding molecular visualization programs for feature parameter extraction [18]. Through this approach, it is possible to better capture the characteristics of speech signals and improve the accuracy of speech recognition and understanding. By analyzing and processing the disassembled speech signals, literature can extract a series of feature parameters that can be used for tasks such as speech recognition, intention understanding, and speech synthesis in intelligent robots. Based on the similarity between feature vectors, the literature adopts a content-based recommendation method to generate a music recommendation list in TopN format, providing users with music recommendations that match their preferences [19]. In the literature, deep learning models play a crucial role [20]. The CRNN model combines the advantages of convolutional neural networks and recurrent neural networks, and can simultaneously capture the temporal and spatial features of audio signals and lyrics. The CRAHRNN model introduces an attention mechanism on the basis of CRNN, allowing the model to focus more on important audio and lyric features. Through these models, audio signals and lyrics can be transformed into high-dimensional feature vectors, thereby measuring the similarity between user interests and music. Selecting the TopN music with the highest similarity as the recommendation result takes into account both the user's personalized preferences and the ability to recommend music that is similar to the music the user has already listened to. The literature proposes a course recommendation model based on dynamic and short-term interests [21]. In this model, GRU (Gated Recurrent Unit) is used to simulate the dynamic changing interests of users, and attention mechanisms are introduced to distinguish the importance of dynamic interests at different times. In course recommendations, user interests may change over time. In order to better capture the dynamic interests of users, the literature uses GRU to model the evolution of user interests at different time steps. GRU is a recursive neural network model with a gating mechanism that can automatically learn and update the user's interest state. By modeling historical behavior sequences, GRU can capture the evolutionary trends of user interests. The attention mechanism can weight dynamic interests at different time steps, allowing important interests to receive more attention. Through this approach, the model can more accurately understand the changes in user interests at different time steps and consider important dynamic interests in the recommendation process. The course recommendation model proposed in the literature comprehensively considers the dynamic interests and short-term interests of users [22]. By using the GRU model to simulate the dynamic evolution of user interests and introducing attention mechanisms to distinguish the importance of interests at different times, this model can better understand the changes in user interests and provide more personalized and accurate course recommendation results during the recommendation process.

## 3. Research on the application of three social robots assisted music course teaching

### 3.1. Social robot teaching

Robot education can stimulate students' innovative thinking and creativity. Robot education can stimulate students' interest in these fields and provide a platform for practical application of the knowledge learned. By participating in robot education, students can enhance their hands-on creativity and problem-solving abilities in the STEM field. In recent years, people have increasingly recognized the adaptability of applying educational robots in educational and teaching environments, and robot education has gradually attracted public attention. By participating in robot education, students can actively participate in practical activities, gain experiences different from theoretical learning, and improve their hands-on ability and teamwork spirit.

Robots have a huge potential impact on education, and as a popular educational method, robot education meets the needs of the development of education in the new era. As digital learning equipment, educational robots involve multiple fields such as science, technology,

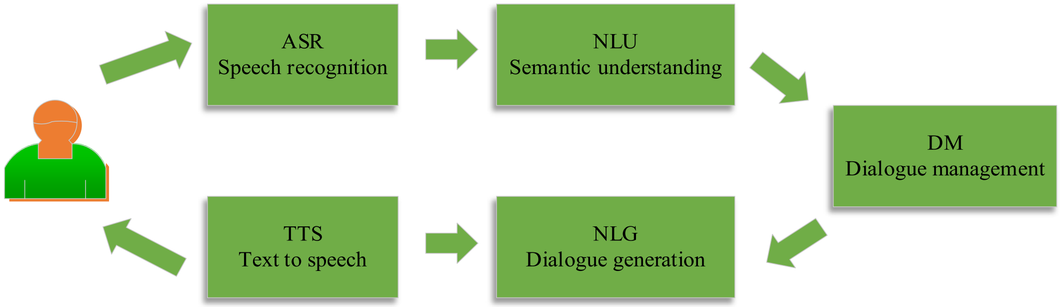

**Fig. 1.** Composition of an intelligent voice interaction system for social robots.

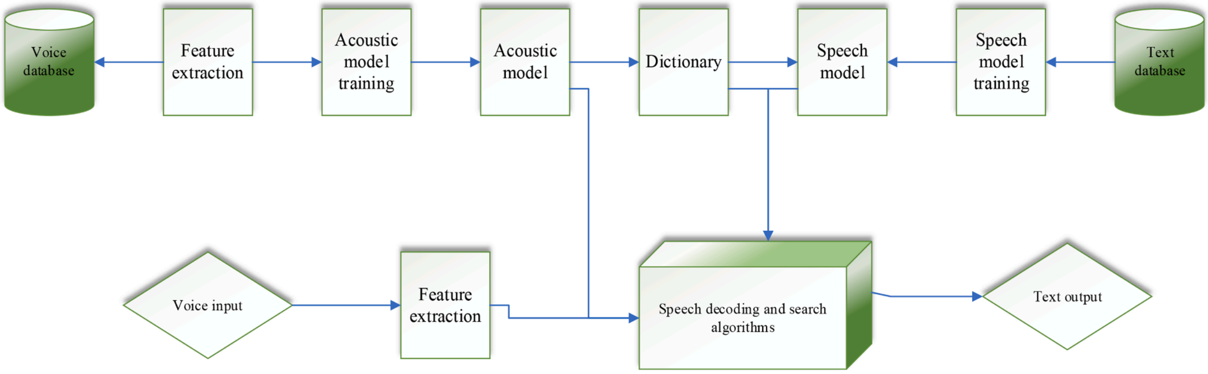

**Fig. 2.** Speech recognition block diagram.

engineering, and mathematics (STEM), and are highly compatible with interdisciplinary learning. They have great potential for promoting the cultivation of learners' hands-on practice, creative problem-solving, and other abilities. Therefore, current research aims to explore the creative potential exhibited by students in learning and using educational robots, and present it. As an innovative educational tool, educational robots provide students with opportunities to participate in practice and exploration. Students need to design, build, and program robots to complete specific tasks, which not only requires them to apply the knowledge they have learned, but also to practice and continuously improve. The experience of teamwork has cultivated students' spirit of collaboration and social skills, laying a solid foundation for their future career development. Robot education can also help students connect their learned knowledge with real-life situations, stimulating their interest in science, technology, and engineering.

### 3.2. *Social robot voice sensing model*

As shown in Fig. 1, the Speech Recognition (ASR) module uses sound signal processing technology and speech recognition algorithms to convert speech signals into text representations for subsequent processing and understanding. The Natural Language Understanding (NLU) module performs semantic parsing and intent recognition on the transformed text instructions or questions, transforming them into machine understandable forms and extracting key information for subsequent processing. The Natural Language Generation (NLG) module is responsible for transforming the system generated answers or feedback into text or speech forms of natural language. It uses natural language generation techniques and models to generate responses or feedback that meet grammar and fluency requirements based on the context of the conversation and user needs. The Text to Speech (TTS) module converts the system generated text into audible speech form.

As shown in Fig. 2, the language model further processes the output of the acoustic model by calculating the probability of possible phrase sequences corresponding to the speech signal.

The goal of speech recognition is to find the text sequence with the highest posterior probability based on the given speech input data. This problem can be represented by formula (1):

$$\begin{aligned} W &= \underset{w}{argmax} P(W|X) \\ &= \underset{w}{argmax} \frac{\mathrm{P(Y|W)P(W)}}{\mathrm{P(Y)}} \\ &\approx \underset{w}{argmax} P(Y|W)P(W) \end{aligned} \tag{1}$$

Among them, W represents a text sequence, and X represents speech input data. To calculate P (W | X), a language model is used to estimate P (W), and an acoustic model is used to estimate P (X | W). The N-gram model can be represented by formula (2):

$$\mathrm{p(W) = p(w_1 w_2 \cdots w_n) = p(w_1)p(w_2|w_1)\cdots p(w_n|w_{1:n-1})} \tag{2}$$

An acoustic model can be a Hidden Markov Model (HMM) or similar, which assigns a corresponding state to each feature vector and estimates the probability of the state sequence. In the N-gram language model, the N-order Markov assumption is made, which means that the probability of each word appearing is only related to the nearest N historical words. This can be expressed using formula (3):

$$\mathrm{p(w_k|w_{1:k-1}) \approx p\left(w_k \middle| w_{k-1-(N-1)k-1}\right)} \tag{3}$$

The feature representation of the hidden layer can be calculated using formula (4):

$$\mathrm{h_t = \sigma\left(W^{(hh)} h_{t-1} + W^{(hx)} x_t\right)} \tag{4}$$

The probability distribution $\mathrm{y_t}$ of the output can be calculated using formula (5):

$$\mathrm{y_t = softmax\left(W^{(S)} h_t\right)} \tag{5}$$

Softmax is a function used to normalize the output, represented by

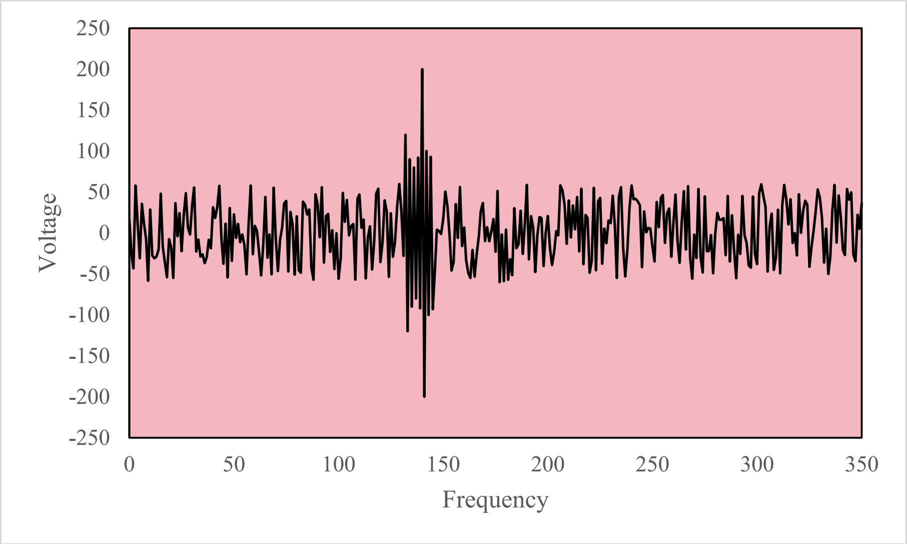

**Fig. 3.** Experimental results of frame processing.

formula (6):

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{6}$$

The cross entropy loss function can be expressed using formula (7):

$$J^{(t)}(\theta) = -\sum_{j=1}^{|V|} y_{t,j} \times \log\left(y_{t,j}\right) \tag{7}$$

The cross entropy loss across the entire corpus can be calculated by averaging the losses of all samples, using formula (8):

$$J = \frac{1}{T}\sum_{t=1}^{T} J^{(t)}(\theta) = -\frac{T}{t=1}\sum_{j=1}^{|||} y_{t,j} \times \log\left(y_{t,j}\right) \tag{8}$$

This article proposes a mathematical model to describe the pulse model of intelligent robot voice signals, using formula (9) to represent:

$$G(z) = \frac{1}{\left(1 - \ell - z^{-1}\right)^2} \tag{9}$$

When transmitting voice signals of intelligent robots in real-time systems, the transfer function of the excitation signal is usually modeled as a pole model. If the speech signal of an intelligent robot contains noise, an autoregressive moving average (AR-MA) model can be used for modeling. The real-time system frequency H(z) can be defined using formula (10):

$$H(z) = \frac{1}{A(z)} \tag{10}$$

When the intelligent robot emits a voice signal, the frequency of the sound wave will undergo resonance, which can be calculated using the voice signal modulation function, i.e. formula (11):

$$V(z) = \frac{A(z)}{1 - Bz - Cz} \tag{11}$$

In order to solve the characteristic parameters of intelligent robot voice signals, the Schuler recursive method was adopted. In the process of Schuler recursion, all the quantities obtained are less than 1, making it very suitable for extracting and calculating feature parameters of speech signals. According to formula (12), use Schuler recursion to calculate the feature parameters of speech signals.

$$Q^m = \sum_{i=0}^{m} a_i r(i-1) \tag{12}$$

According to the properties of $Q^m$ and the positive correlation theorem, it can be proven that the feature coefficients $K^m$ and $Q^m$ of speech signals satisfy formula (13):

$$K^m = -\frac{Q^{(m-1)}}{Q^{(m+1)}} \tag{13}$$

Through such extraction and calculation, the characteristics of speech signals can be further analyzed, speech recognition algorithms can be improved, the quality of speech interaction can be improved, and the performance of intelligent robots can be enhanced. Frame addition processing refers to the process of overlapping and superimposing consecutive speech frames in speech signal processing to obtain more data information. Frame addition processing is usually applied to the feature extraction process of speech signals, such as MFCC feature extraction in speech recognition. By overlapping and superimposing consecutive speech frames, the sample size of feature data can be increased, and the stability and accuracy of features can be improved. In frame addition processing, a fixed window size is usually used for frame segmentation, and an overlap value is set to determine the degree of overlap between adjacent frames. The commonly used window functions include Hamming window, Haining window, etc. The experiment used frame addition processing, and the results are shown in Fig. 3.

As shown in Fig. 3, frame addition processing can effectively complete data processing. By using the windowing processing method proposed in the article, it can be ensured that the ideal data window is used to offset the external impact response during the frame addition process, thereby ensuring the extraction effect of feature parameters. Window processing is a crucial step in the framing process, aimed at reducing spectrum leakage caused by frame truncation and increasing the smoothness of the spectrum. The commonly used window functions

**Table 1**
Experimental results of feature parameter extraction time/min.

| Number of experiments | Method 1 | Method 2 | Method of this article |
|---|---|---|---|
| 1 | 1.242 | 2.027 | 0.251 |
| 2 | 1.481 | 2.196 | 0.228 |
| 3 | 1.415 | 1.904 | 0.278 |
| 4 | 1.248 | 1.422 | 0.361 |
| 5 | 1.408 | 1.259 | 0.212 |
| 6 | 1.257 | 1.657 | 0.180 |

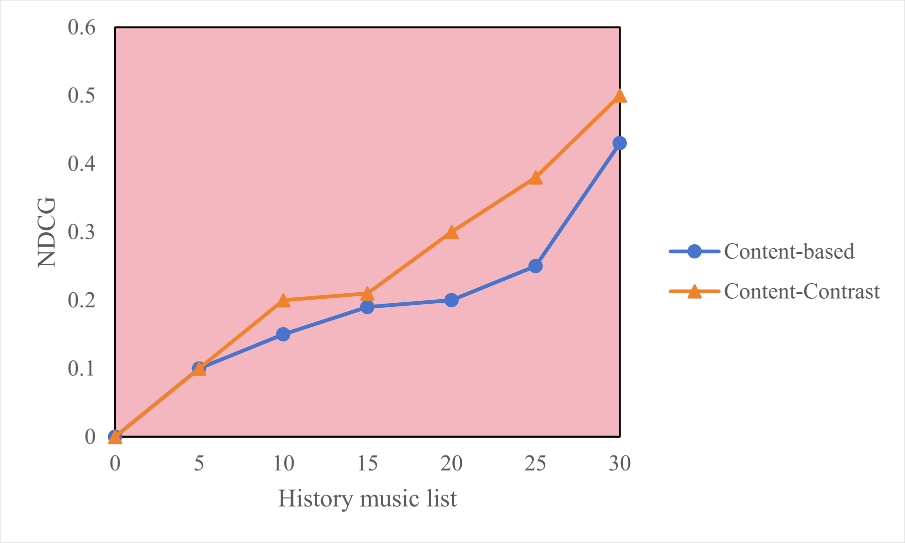

**Fig. 4.** The impact of the number of historical songs listened to on algorithm performance.

include Hamming window, Haining window, etc. By using ideal data windows to counteract external shock responses, windowing can effectively reduce spectrum leakage caused by frame truncation. By ensuring the extraction effect, more accurate feature parameters of speech signals can be obtained, providing a more accurate data foundation for subsequent speech processing and recognition tasks.

The experimental results of feature parameter extraction time are shown in Table 1.

Based on the experimental results and comparative analysis in Table 1, this method mainly uses two feature parameters: cepstral coefficient and linear prediction coefficient. These two coefficients mainly simulate the sound generation device of intelligent robots, without considering the auditory system of intelligent robots. These feature parameters can provide good speech signal recognition ability, with small computational complexity and easy implementation. Among them, the cepstral coefficient is calculated by performing Fourier transform and logarithmic compression on the speech signal, which can reflect the characteristics of the speech signal in the frequency domain and has good discriminability. The linear prediction coefficient is obtained by linearly predicting the speech signal, which can reflect the time-domain characteristics of the speech signal. The combination of these two coefficients can comprehensively describe the characteristics of the speech signal, thereby achieving effective extraction and analysis of the speech signal of intelligent robots. Therefore, by using cepstral coefficients and linear prediction coefficients as feature parameters, this method can extract speech signal features with obvious distinctiveness and independence in a short period of time. This method has good performance in intelligent robot speech signal processing, and has low computational complexity and is easy to implement.

## 4. A deep learning recommendation algorithm model for music course resources

### *4.1. Deep learning techniques*

Big data refers to a vast collection of data, and deep learning models can use big data for training to learn better data representations. Deep learning models can automatically extract features from large amounts of data and establish complex patterns and correlations. In order to complete complex computational tasks in the model, such as matrix operations and gradient optimization, the Graphics Processing Unit (GPU) provides powerful parallel computing capabilities, making the training process of deep learning models more efficient and fast. Deep learning has a high demand for data, as it requires a large amount of data to fully support its rich parameterization. However, in some fields such as language or vision, label data is relatively scarce, making it relatively easy to collect large amounts of data in the context of recommendation system research.

### *4.2. Music feature extraction*

The basic steps of building a deep learning model for music recommendation include preparation work and formal training process. In the preparation work before training, the first step is to vectorize the audio signal and lyrics information. Audio processing techniques such as Mel spectral feature extraction can be used to convert audio signals into feature vectors, while natural language processing techniques can be used to convert lyric information into text feature vectors. The resulting audio and lyric vectorized data will serve as input for deep learning models. During the formal training process, deep learning models can be used to extract audio and lyric features, which can be achieved by constructing neural network models suitable for audio and text processing. For example, convolutional neural networks (CNNs) can be used to process audio features, or recurrent neural networks (RNNs) can be used to process lyrics features. By training on large-scale datasets, the model can learn complex associations between music and lyrics. After obtaining the trained model, the similarity between the user feature vector (user interest feature) and the item feature vector (music feature vector) can be calculated. A common method is to use metrics such as cosine similarity or Euclidean distance to measure the degree of similarity between them.

Pre emphasis is achieved by applying a high pass filter to the audio signal, which enhances the amplitude of the high-frequency portion.

$$y(n) = x(n) - 0.97x(n-1) \tag{14}$$

Pre emphasis can reduce possible abrupt changes at frame boundaries and help improve the performance of subsequent audio analysis algorithms. The purpose of framing is to analyze audio signals in a short period of time to capture short-term changes in audio signals over time. The number of frames can be determined based on formulas (15) and (16).

$$N_w = 10^{-3}(fs.T_w) \tag{15}$$

$$N_s = 10^{-3}(fs.T_s) \tag{16}$$

By dividing the audio signal into multiple frames through framing

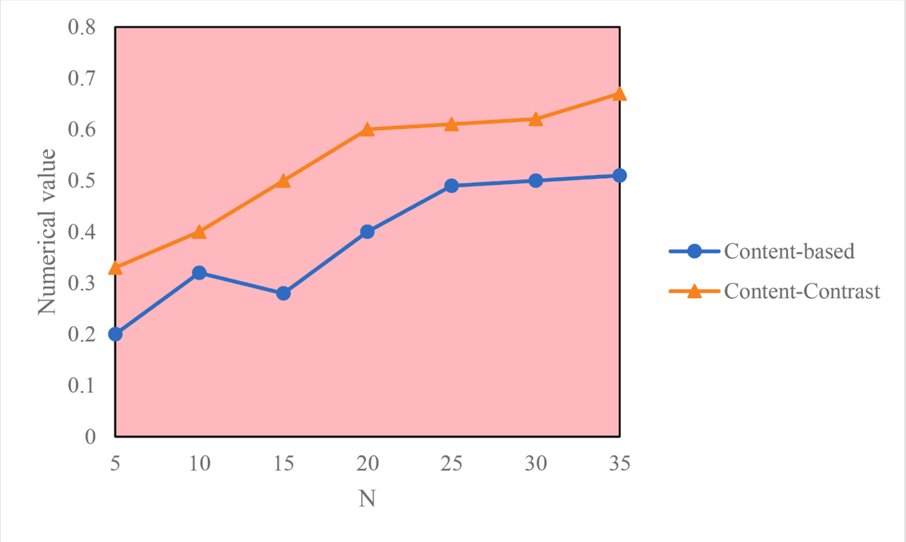

**Fig. 5.** Recall rate results.

operations, independent spectral analysis and feature extraction can be performed on each frame to better capture the temporal and frequency features of the audio signal. Framing operations are widely used in tasks such as audio processing and speech recognition. The purpose of adding windows is to smooth the signal. By convolving the window function and input signal, signal leakage can be reduced and the smoothing effect can be achieved. This article uses Hamming windows to achieve smoothing., See the following formula:

$$\mathrm{w(n)} = 0.54 - (0.46\cos(2\pi\mathrm{n}/\mathrm{N}-1));\ \mathrm{for\,n} = 0, 1, 2, \cdots, \mathrm{N}-1 \tag{17}$$

Spectral analysis can understand the frequency components and energy distribution of signals. By analyzing the spectrum, the spectral features of audio signals, such as spectrum envelope and flatness, can be extracted. After FFT processing, the amplitude of the FFT result will be adjusted in the Mel frequency scale. The Mel frequency scale is a specific frequency scale that converts linear frequencies below 1 kHz and logarithmic frequencies above 1 kHz. The Mel scale can be obtained by formula (18):

$$\mathrm{Mel(f)} = 1125\log_{10}\left(1+\frac{\mathrm{f}}{700}\right) \tag{18}$$

Here, n is defined as the word count of the longest lyric among all lyrics; Then lM can be represented by formula (19):

$$\mathrm{lM}_{1:n} = \mathrm{lwV}_{ij1} \oplus \mathrm{lwV}_{ij2} \oplus \cdots \oplus \mathrm{lwV}_{ijn} \tag{19}$$

Each word can generate a feature value fcx through filter bank F, and the process is as follows:

$$\mathrm{fc}_{\mathrm{x}} = \mathrm{f}(\mathrm{F}\cdot\mathrm{lM}_{\mathrm{x-h+1:x}}+\mathrm{b}) \tag{20}$$

This article investigates the impact of user's historical listening volume on content based and music recommendation algorithms, and compares them using the NDCG metric, as shown in Fig. 4.

As shown in Fig. 4, when the user listens to a small amount of music, the two algorithms are very close or even overlap. This is because when the user listens to a small amount, their personality characteristics are not obvious enough, and the two algorithms cannot clearly distinguish the user's preferences. But when users listen to more than 10 pieces of music, the music recommendation algorithm in this article gradually becomes stronger than content-based recommendation algorithms, and has maintained a leading position since then. This result can be explained as when the number of users listening is small, the algorithm mainly relies on the music content itself for recommendation. Due to the lack of clear individual characteristics of users, both algorithms find it difficult to accurately capture their preferences. But as the number of users listening increases, the optimization of this article can better discover their personalized preferences and provide more accurate recommendation results by comparing their historical listening data with other users. Therefore, when the number of users listening is large, the music recommendation algorithm in this article can better meet the preferences of users, and performs better compared to content-based algorithms. Therefore, when designing recommendation algorithms, one should consider the user's historical listening volume and flexibly adjust the algorithm's strategy. For users with fewer listening numbers, more content based recommendation algorithms can be adopted; For users with a large listening volume, the music recommendation algorithm in this article can provide more accurate recommendation results.

Fig. 5 shows the recall rates of content-based recommendation algorithms and music recommendation algorithms in different TopN recommendations. From Fig. 5, it can be observed that the recall rate of our algorithm is significantly better than the original content-based recommendation algorithm.

### 4.3. *Conversation interest extraction*

The framework of this article aims to generate a mixed preference representation of users by integrating their dynamic and short-term preferences. This framework mainly consists of three levels. The user's historical session can be viewed as a temporal task, where there is a certain sequential relationship between each session and the preceding and following sessions. The GRU (Gated Recurrent Unit) model can solve the problem of long-term dependencies, which means it can capture dependencies over longer time intervals.

In the modeling process, historical sessions are used as input sequences and fed into the gated loop unit. Formulas (21) to (24) are used to calculate the hidden state and output.

$$z_t = \sigma(W_z\cdot[h_{t-1}, I_t]) \tag{21}$$

$$r_t = \sigma(W_r\cdot[h_{t-1}, I_t]) \tag{22}$$

$$\widetilde{h}_t = tanh(W\cdot[r_t{}^*h_{t-1}, I_t]) \tag{23}$$

$$h_t = (1-z_t)^*h_{t-1} + z_t{}^*\widetilde{h}_t \tag{24}$$

By embedding the output of the session interest extraction layer as the input of the dynamic interest representation layer, the HAGRU model is used to model the changes in user interests in different

**Table 2**
Comparison Table of Baseline Effects.

| option | AUC | Precision@5 | Precision@10 | Precision@50 | Precision@100 |
|---|---|---|---|---|---|
| SLIM | 0.752 | 0.083 | 0.062 | 0.026 | 0.015 |
| BPR | 0.765 | 0.104 | 0.083 | 0.043 | 0.029 |
| FPMC | 0.766 | 0.112 | 0.942 | 0.046 | 0.020 |
| NCF | 0.799 | 0.142 | 0.111 | 0.057 | 0.041 |
| DeepFM | 0.800 | 0.132 | 0.115 | 0.058 | 0.039 |
| SHAN | 0.751 | 0.140 | 0.116 | 0.068 | 0.047 |
| HAGRU | 0.837 | 0.156 | 0.135 | 0.083 | 0.062 |

**Table 3**
Comparison of variant effects.

| option | AUC | Precision@10 |
|---|---|---|
| HAGRU | 0.823 | 0.157 |
| HAGRU-A | 0.814 | 0.130 |
| HAGRU-B | 0.799 | 0.128 |

**Table 4**
The influence of different regularization parameters under AUC.

| $\lambda_{uc}/\lambda_a$ | 0 | 1 | 5 |
|---|---|---|---|
| 0.01 | 0.770 | 0.827 | 0.776 |
| 0.001 | 0.759 | 0.751 | 0.760 |
| 0.0001 | 0.766 | 0.789 | 0.763 |

historical sessions. This can better understand and represent the dynamic interest changes of users, providing strong support for personalized recommendations, behavior analysis and other tasks. Formula (25) is used to calculate the attention weight for each time step t, and formula (26) is used to normalize the attention weight to meet the requirements of probability distribution.

$$f_{2t} = \Phi(W_2 h_t + b_2) \tag{25}$$

$$\alpha_t = \frac{exp(u^\top f_{2t})}{\sum_{p \in L^u_{t-1}} exp\left(u^\top f_{2p}\right)} \tag{26}$$

Finally, by weighting the hidden state ht with the corresponding attention weight αt using formula (27), a vector representation representing the user's dynamic preferences is obtained, which takes into account the importance of the user's dynamic interest at different time steps.

$$u_t^{\text{dynamic}} = \sum_{t \in L^u_{t-1}} \alpha_t h_t \tag{27}$$

By introducing attention mechanisms, it is possible to more accurately capture the dynamic interests and preferences of users at different times, and provide more accurate information for subsequent personalized recommendations, interest analysis, and other tasks. The attention mechanism enables the model to model the dynamic changes in user interests, thereby improving the model's expressive ability when processing temporal data.

The HAGRU model utilizes the first layer attention mechanism to remove course noise and extract user interests in conversations. This article investigates the performance comparison of the HAGRU model with all baseline methods, as shown in Table 2.

In order to evaluate the impact of session interest extraction layer and dynamic interest representation layer on the recommendation model, variant experiments were conducted. The experimental results are shown in Table 3:

As shown in Table 3, it is observed that HAGRU-A is present in both AUC and Precision@10 The performance is superior to HAGRU-B, indicating that capturing the dynamic changes of users is more important for recommendation tasks compared to ignoring noisy courses. The experimental results show that obtaining high-quality user interest representations at the bottom layer contributes more to improving the accuracy of recommendation models than deep learning models.

According to the data in Table 4, it was observed that different regularization parameters have an impact on AUC.

Table 4 shows that appropriate selection of regularization parameters can improve the performance of the model in recommendation tasks. In this experiment, the best results were achieved with a regularization parameter of 0.01, while smaller regularization parameters (0.001 and 0.0001) resulted in poorer recommendation performance.

## 5. Conclusion

Currently, the field of music education is gradually exploring and applying voice sensing and deep learning algorithms, using social robots as auxiliary teaching methods, which has brought revolutionary changes. This new teaching method injects new vitality and possibilities into traditional music education. The application of social robots designed in this article in music education enables students to learn and interact with robots through intelligent interaction, thereby obtaining personalized guidance on music skills. Through voice sensing technology, social robots can understand the needs and problems of students, injecting new vitality into music education, breaking the limitations of traditional teaching, and expanding the possibilities of teaching. Through interactive learning with social robots, students can better understand and master music knowledge, stimulate learning interest, improve learning efficiency, and thus bring richer and more diverse teaching experiences to music education. Therefore, the application of social robots in the field of music education has great potential and significance, which will bring more innovation and possibilities for future music education, improve teaching quality, and promote the progress of education methods. This new model that combines technology and education will undoubtedly become an emerging development direction in the field of education, providing students with more attractive and effective learning experiences.

## Funding

This study and all authors have received no funding.

## CRediT authorship contribution statement

**Xiao Dan:** Writing – original draft, Methodology, Formal analysis.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## References

[1] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka, Social robots for education: a review, Sci. Rob. 3 (21) (2018) eaat5954.

[2] E.P. Asmus, Motivation in music teaching and learning, Visions Res. Music Educ. 16 (5) (2021) 31.

[3] S.M. Anzalone, S. Boucenna, S. Ivaldi, M. Chetouani, Evaluating the engagement with social robots, Int. J. Soc. Robot. 7 (2015) 465–478.

[4] A. Shrestha, A. Mahmood, Review of deep learning algorithms and architectures, IEEE Access 7 (2019) 53040–53065.

[5] R. Van den Berghe, J. Verhagen, O. Oudgenoeg-Paz, S. Van der Ven, P. Leseman, Social robots for language learning: A review, Rev. Educ. Res. 89 (2) (2019) 259–295.

[6] O. Nocentini, L. Fiorini, G. Acerbi, A. Sorrentino, G. Mancioppi, F. Cavallo, A survey of behavioral models for social robots, Robotics 8 (3) (2019) 54.

[7] A.A. Scoglio, E.D. Reilly, J.A. Gorman, C.E. Drebing, Use of social robots in mental health and well-being research: systematic review, J. Med. Internet Res. 21 (7) (2019) e13322.

[8] C. Giang, A. Piatti, F. Mondada, Heuristics for the development and evaluation of educational robotics systems, IEEE Trans. Educ. 62 (4) (2019) 278–287.

[9] M. Benedek, N. Nordtvedt, E. Jauk, C. Koschmieder, J. Pretsch, G. Krammer, A. C. Neubauer, Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers, Think. Skills Creat. 21 (2016) 75–84.

[10] S. Humphrey-Murto, M. De Wit, The Delphi method—more research please, J. Clin. Epidemiol. 106 (2019) 136–139.

[11] L. Zhao, L. Li, Y. Wu, Research on the coupling coordination of a sea–land system based on an integrated approach and new evaluation index system: A case study in Hainan Province, China. Sustainability 9 (5) (2017) 859.

[12] M.S. Islam, M.M. Rahman, G. Muhammad, M.S. Hossain, Design of a social robot interact with artificial intelligence by versatile control systems, IEEE Sens. J. 22 (18) (2021) 17542–17549.

[13] V.Z. Kepuska, M.M. Eljhani, B.H. Hight, Voice activity detector of Wake-Up-Word speech recognition system design on FPGA, J. Eng. Res. Appl. 4 (12) (2014) 160–168.

[14] M. Peng, Y. Qin, C. Tang, X. Deng, An e-commerce customer service robot based on intention recognition model, J. Electron. Commerce Organizations (JECO) 14 (1) (2016) 34–44.

[15] J. Kong, J. Kim, J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, Adv. Neural Inf. Proces. Syst. 33 (2020) 17022–17033.

[16] A.B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev, T.K. Whangbo, Improved feature parameter extraction from speech signals using machine learning algorithm, Sensors 22 (21) (2022) 8122.

[17] A. Davletcharova, S. Sugathan, B. Abraham, A.P. James, Detection and analysis of emotion from speech signals, Procedia Comput. Sci. 58 (2015) 91–96.

[18] C.R. Terrell, L.L. Listenberger, Using molecular visualization to explore protein structure and function and enhance student facility with computational tools, Biochem. Mol. Biol. Educ. 45 (4) (2017) 318–328.

[19] A.H. Nabizadeh, J.P. Leal, H.N. Rafsanjani, R.R. Shah, Learning path personalization and recommendation methods: A survey of the state-of-the-art, Expert Syst. Appl. 159 (2020) 113596.

[20] C. Bhatt, I. Kumar, V. Vijayakumar, K.U. Singh, A. Kumar, The state of the art of deep learning models in medical science and their challenges, Multimedia Syst. 27 (4) (2021) 599–613.

[21] V.A. Nguyen, H.H. Nguyen, D.L. Nguyen, M.D. Le, A course recommendation model for students based on learning outcome, Educ. Inf. Technol. 26 (2021) 5389–5415.

[22] G. Xu, G. Jia, L. Shi, Z. Zhang, Personalized course recommendation system fusing with knowledge graph and collaborative filtering, Comput. Intell. Neurosci. 2021 (2021) 1–8.