Research on the Design of Intelligent Voice Interaction System Based on Affective Computing

Jialin Sun

Shandong University of Science and Technology, Shandong, China

Received: 12 Jul 2025
Revised: 15 Jul 2025
Accepted: 18 Jul 2025
Published: 30 Jul 2025

Abstract: With the rapid development of intelligent voice technology, traditional voice interaction systems have gradually been unable to meet users' needs for emotional interaction. To improve the naturalness and humanity of human-computer interaction, this study proposed and designed an intelligent voice interaction system based on emotional computing. The system accurately analyzes the user's emotional state through emotion recognition technology and combines speech synthesis technology to generate voice feedback that meets the user's emotional needs. The study first discussed the relevant technologies of emotion computing and voice interaction in detail, designed the system architecture, and implemented emotion recognition and speech generation algorithms. Through experimental evaluation and user feedback analysis, the superiority of the system in emotion recognition accuracy, voice interaction effect, and user experience were verified. Compared with traditional voice interaction systems, the system of this study significantly improved the user's emotional resonance and interaction satisfaction, reflecting the practical application value of emotion computing in voice interaction. Finally, the study discussed the limitations and future development directions of the system and proposed the possibilities and challenges for the further development of emotion computing and voice interaction.

Keywords: Affective computing; Intelligent voice interaction; Emotion recognition; Speech synthesis; User experience; Human-computer interaction

1 INTRODUCTION

With the continuous development of artificial intelligence (AI), affective computing, as a key research area, has gradually garnered widespread attention from both academia and industry. Affective computing refers to the use of computer systems to identify, analyze, understand, and generate human emotions. It has applications in fields such as speech recognition, natural language processing, and computer vision [1]. Research on affective computing began in the 1990s, initially focusing on emotion recognition. With the advancement of technology, emotion generation and emotional feedback have gradually become key areas of research. In the field of voice interaction, in particular, the application of affective computing is not limited to improving speech recognition accuracy but also aims to enhance the naturalness of human-computer interaction and the richness of emotional expression [2].

As a key component of AI technology, intelligent voice interaction systems have made significant progress in recent years. Traditional voice interaction systems primarily focus on speech recognition and command execution, prioritizing functionality and efficiency. However, with technological advancements and increasing user demands, relying solely on mechanical voice interaction can no longer meet users' demands for a natural and personalized interactive experience [3]. Currently, challenges facing intelligent voice interaction systems include insufficient emotional understanding, monotonous speech output, and unnatural interaction. Therefore, the introduction of affective computing provides new ideas for the upgrade and optimization of voice interaction systems, which can make voice interaction more humane and enhance its emotional level and feedback effect.

The primary objective of this research is to design and optimize an intelligent voice interaction system based on affective computing technology, enhancing its capabilities in emotion recognition, generation, and feedback, further improving the system's naturalness and user experience. In traditional voice interaction systems, speech delivery is often mechanical and monotonous, making it difficult to establish a deep emotional connection with users. The introduction of affective computing, however, allows the system to dynamically adjust based on the user's emotional state, making voice interaction more humane and personalized, thereby increasing user satisfaction and interaction effectiveness [4].

Through this design, the goal is to not only enable the voice interaction system to recognize and understand user emotions, but also to generate speech output that aligns with user emotional needs. This will enable the system to operate in more complex and realistic interaction scenarios, resulting in a smoother and more comfortable user experience. Furthermore, this research aims to provide theoretical foundations and technical support for the further development of intelligent voice systems and promote the widespread application of affective computing in the field of voice interaction.

The core approach of this research is to integrate affective computing technology into intelligent voice interaction systems, combining it with emotion recognition and generation algorithms to design a new voice interaction model [5]. Specifically, the research will begin with the collection and analysis of emotional data. Leveraging machine learning and deep learning techniques, the research will develop an emotion recognition module. This module analyzes and identifies the user's emotional state using multi-dimensional features in speech signals, such as intonation, voice tone, speaking rate, and volume. Simultaneously, the research will also design an emotion generation algorithm to ensure the system can generate appropriate speech output based on the user's emotional feedback, enhancing the naturalness and affinity of human-computer interaction [6].

The research's technical approach consists of three phases: First, the collection and preprocessing of emotional data to create a multi-dimensional emotion dataset; second, the construction and optimization of emotion recognition and generation models, using deep learning algorithms to train and optimize emotion features; and third, system implementation and experimental testing to evaluate the effectiveness of affective computing in improving the performance of intelligent voice interaction systems [7]. During the research process, experiments will be conducted using existing speech datasets and validated in real-world scenarios, ultimately realizing the effective application of affective computing in voice interaction.

This paper is divided into seven Sections. Section 1 is the introduction, which mainly introduces the background, purpose and significance of the research, the research methods and technical approach, and the structure of the paper. Section 2 will provide a detailed analysis of the basic theory of affective computing, expounding on the definition and development history of affective computing, and the technical key points of emotion recognition and generation. Section 3 will review the relevant research on intelligent voice interaction systems, analyze the status and challenges of voice interaction systems, and explore the application potential of affective computing in them. Section 4 will provide a detailed introduction to the design and implementation of an intelligent voice interaction system based on affective computing, focusing on the system architecture, the functional design of each module, and the integration of affective computing technology. Section 5 will provide a detailed explanation of the experimental part, including the system implementation process, experimental environment and testing methods, as well as the evaluation and analysis of experimental results. Section 6 will summarize and discuss the research results, analyze the shortcomings and limitations of the research, and propose future research directions. Section 7 is the conclusion, summarizing the main research results of the paper and looking forward to the application prospects of affective computing in intelligent voice interaction systems.

2 FOUNDATIONS OF AFFECTIVE COMPUTING

2.1 Concept and Development of Affective Computing

Affective computing refers to the technology used to identify, understand, generate, and express human emotions through computer systems, aiming to enable computers to simulate or understand human emotions. Research on affective computing originated in the 1990s, initially focusing on emotion recognition, that is, identifying a person's emotional state through speech, facial expressions, or other physiological signals [8]. With the continuous advancement of technology, affective computing has expanded beyond emotion recognition to include emotion generation and feedback. The technologies involved have also expanded from artificial intelligence and pattern recognition to include deep learning, natural language processing (NLP), computer vision, and other fields.

Affective computing is widely used in various industries, including intelligent customer service, virtual assistants, smart healthcare, education, and entertainment. In these fields, the introduction of affective computing can not only improve the quality of interaction between systems and users but also enhance the user's emotional experience to a certain extent. Such as, in intelligent voice interaction, affective computing enables the system to recognize changes in user emotions and adjust the tone and intonation of the voice, making the voice interaction more humane and emotionally rich [9]. With the continuous advancement of affective computing technology, it has gradually become a key technology for improving the intelligent interaction experience, playing an increasingly important role in emotionally rich voice interaction.

2.2 Emotion Recognition Technology

Emotion recognition is one of the core technologies in affective computing. Its goal is to identify an individual's emotional state by analyzing multiple signals (Such as speech, facial expressions, and body posture). In voice interaction, voice emotion recognition has become the most common method for emotion recognition [10]. Speech signals contain rich emotional information, such as intonation, speaking rate, and volume, which can reflect the speaker's emotional state. The implementation of speech emotion recognition technology relies on the extraction and analysis of these acoustic features. Machine learning or deep learning algorithms are typically trained on large amounts of speech data to achieve accurate emotion classification.

Common techniques for speech emotion analysis include acoustic feature-based emotion analysis, emotion analysis based on sentiment lexicons, and deep learning methods such as deep neural network (DNN) and convolutional neural network (CNN). These algorithms can extract emotional features from audio signals and combine them with contextual information to perform emotion classification. Currently, several public datasets are available for emotion recognition research, such as Emo-DB and RAVDESS [11]. These datasets contain speech samples in various emotional states and can be used to train and evaluate emotion recognition models. At the same time, the evaluation standards for emotion recognition are constantly being improved. Common evaluation indicators include accuracy, recall rate, F1 value, etc.

2.3 Emotion Generation and Expression

Emotion generation is another key component of affective computing, enabling computer systems to generate appropriate emotional feedback based on emotional states. In voice interaction systems, emotion generation typically manifests itself as speech output, using speech synthesis technology to generate emotionally charged speech. To make speech more natural and emotional, emotion generation models adjust various speech features, such as speech rate, intonation, pauses, and accents, based on the recognized emotional state.

Natural language generation (NLG) technology plays a particularly important role in emotion generation. In traditional speech synthesis systems, speech generation usually relies on pre-set rules and templates. In systems based on emotional computing, NLG technology can generate natural language content that meets emotional needs based on the input emotional information. Through technologies such as deep learning and generative adversarial networks (GAN), emotion generation models can gradually achieve more flexible and expressive speech output. The successful implementation of emotion generation not only makes speech synthesis closer to the expression of natural human language but also can adjust speech output according to user needs in different emotional scenarios, thereby improving the interactive experience.

2.4 Application of affective computing in voice interaction

The application of affective computing in voice interaction has greatly enhanced the naturalness and emotional experience of human-computer interaction. Traditional voice interaction systems primarily focus on voice recognition and command execution, often ignoring the user's emotional needs, making the interaction appear stilted and impersonal. The introduction of affective computing, however, enables the system to identify the user's emotional state and adjust the content and method of voice output based on these changes, making the interaction livelier and more engaging.

The application of affective computing in voice interaction is primarily reflected in emotion recognition. Through emotion recognition, voice interaction systems can identify user emotions such as joy, anger, and anxiety and adjust their responses accordingly. Such as, if a user expresses anxiety while interacting with a voice assistant, the system can respond with a calmer, more soothing tone, alleviating the user's emotions. Furthermore, affective computing can also enhance the emotional expression of speech during speech generation through emotion generation models, making the speech output more responsive to the user's emotional needs.

Affective computing is closely related to user experience. By enhancing the emotional depth and emotional relevance of interactions, it significantly increases user engagement and satisfaction with voice interaction. When users experience emotional responses from the system during voice interactions, they develop a greater sense of engagement and trust. In applications like intelligent customer service and virtual assistants, the inclusion of affective computing can not only optimize the user experience but also, to a certain extent, enhance user loyalty to the system. Therefore, affective computing plays an indispensable role in enhancing user experience and strengthening user emotional connections.

3 OVERVIEW OF INTELLIGENT VOICE INTERACTION SYSTEM

3.1 Model Basic Structure of Intelligent Voice Interaction System

Intelligent voice interaction systems are complex systems that combine speech recognition, speech synthesis, and NLP technologies to enable natural and smooth voice interaction between humans and computers. Speech recognition technology is the foundation of intelligent voice interaction systems, converting user voice signals into text for analysis and processing. The core technologies of speech recognition include acoustic models, language models, and decoders. By extracting and matching features from speech signals, they identify words and sentences in the speech, thereby understanding user commands or questions.

Speech synthesis technology converts system responses into speech, providing users with audible feedback. Traditional speech synthesis relies on rules and templates, while modern synthesis technologies (Such as deep learning-based WaveNet) make speech synthesis more natural and fluent, even capable of simulating different emotional tones and intonations. The key to speech synthesis lies in generating speech output based on text content while ensuring naturalness and emotional expression.

NLP is the "brain" of intelligent voice interaction systems, responsible for in-depth analysis and understanding of textual information. NLP technology encompasses multiple aspects, including semantic understanding, syntactic analysis, and sentiment analysis. It enables systems to not only recognize literal information in speech but also understand user intent and emotions, enabling more intelligent and personalized responses. Through semantic understanding and contextual analysis, NLP technology helps systems respond accurately in complex scenarios, making voice interactions more humane.

3.2 Application Areas of Intelligent Voice Interaction

Due to their convenience and efficiency, intelligent voice interaction systems have been widely adopted across multiple industries, driving the development of smart homes, personal assistants, customer service, health management, education, and entertainment. In the smart home and personal assistant sectors, voice interaction provides users with a convenient way to control smart devices such as lighting, air conditioning, and audio systems through voice commands. Virtual personal assistants such as Siri and Alexa can understand and implement users' daily needs, while also providing personalized recommendations and services through voice, improving users' quality of life.

In the customer service and health management sectors, intelligent voice interaction systems can effectively replace human customer service representatives, provide 24/7 service and reduce business operating costs. Specifically in the health management sector, voice assistants can provide personalized health advice, medication reminders, emotional counseling, and other services tailored to user needs, helping users better manage their health. In the education and entertainment sectors, the application of intelligent voice systems not only enhances the learning experience but also provides users with richer entertainment content. Educational voice assistants can provide personalized learning resources through interaction with students, while voice assistants in the entertainment field can recommend movies, TV shows, music, and other content based on users' interests, creating an immersive entertainment experience.

3.2 Analysis of the Emotion Processing Capabilities of Existing Voice Interaction Systems

While existing intelligent voice interaction systems have made significant progress in speech recognition and text generation, the introduction of affective computing is still in its early stages. Currently, the application of affective computing in intelligent voice systems is limited, primarily focused on emotion recognition and simple emotional feedback. Most voice interaction systems have relatively basic emotional processing capabilities, typically limited to assessing a user's emotional state based on voice characteristics such as intonation and speech rate. They also often struggle to recognize and respond to complex emotions. This can lead to inappropriate responses when users experience significant emotional fluctuations or express dissatisfaction, failing to effectively soothe them.

Despite this, some leading intelligent voice systems have begun experimenting with incorporating affective computing to enhance the user experience. Such as, some intelligent voice assistants can analyze emotional cues in the user's voice and adjust the tone of voice output to make it more calming or enthusiastic, thereby creating a more humane interactive experience. However, these affective computing technologies still face numerous challenges in practical applications. First, the accuracy of emotion recognition still needs to be improved, especially in multimodal emotion recognition, where the integration of speech signals with other emotional signals such as facial expressions and body language has not yet achieved ideal results. Second, emotion generation technology is still immature. While current speech synthesis technologies can mostly generate emotional speech, they often fail to accurately capture the complex emotional needs of users.

Therefore, existing intelligent voice systems have certain limitations in terms of emotional feedback, particularly in terms of the diversity of emotional expression and the adaptability of emotional responses. To better address users' emotional needs, future intelligent voice interaction systems will need to further optimize emotion recognition and generation technologies, enhance the system's ability to understand and express emotions, and thereby enhance the naturalness of human-computer interaction and the user's emotional experience.

4 DESIGN OF INTELLIGENT VOICE INTERACTION SYSTEM BASED ON AFFECTIVE COMPUTING

4.1 System Requirements Analysis

The design of an intelligent voice interaction system based on affective computing requires a thorough analysis of user needs and system objectives. Users need primarily focus on enhancing the naturalness of voice interactions, emotional responses, and personalized services. With the prevalence of smart devices, users increasingly expect voice interaction systems to not only understand their commands but also recognize and respond to their emotional changes. Therefore, the system's objectives must be clearly defined: using affective computing technology, the voice interaction system should be able to more accurately identify the user's emotional state and dynamically adjust speech output based on emotional information, thereby enhancing user engagement and emotional satisfaction.

Functionally, the system should be able to collect user voice signals in real time, rapidly analyze their emotional state, and generate speech feedback tailored to their emotional needs through a speech synthesis module. Furthermore, the system should be highly flexible and adaptable, providing a personalized interactive experience tailored to the user's varying emotional states. Non-functional requirements include system efficiency and real-time performance, ensuring smooth emotion recognition and generation in practical applications without delays or identifications. Furthermore, system stability and fault tolerance must be ensured.

4.2 System Architecture Design

The system's overall architecture should encompass modules such as affective computing, speech recognition, speech synthesis, and NLP, ensuring effective collaboration between these modules. The core goal of this architecture is to seamlessly integrate these technical modules, enabling affective computing to play a key role in voice interaction. The system design begins with building an efficient speech recognition module to convert user voice signals into text, which is then analyzed semantically by the NLP module. By combining contextual analysis with emotion recognition algorithms, the system can understand the user's emotional state during interaction, providing a basis for subsequent emotional feedback.

The core of the system architecture lies in the integration of the affective computing module. In combining speech recognition and emotion recognition modules, affective computing not only identifies emotional features in speech but also understands user intent. The affective computing module uses deep learning algorithms to analyze acoustic features such as pitch, rate, and pauses in speech to infer the user's emotional state. Combined with the semantic information provided by the NLP module, the system can adjust subsequent speech generation, and interaction strategies based on the emotion recognition results, ensuring that the speech output is more aligned with the user's emotional needs. This modular design enables the system to not only execute simple voice commands but also enhances the quality of interaction through emotional feedback.

4.3 Emotion Recognition and Generation Algorithm Design

Emotion recognition and generation algorithms are core technologies in intelligent voice interaction systems based on affective computing, directly determining the system's ability to process emotions and the effectiveness of its interactions. The collection and preprocessing of emotional data is the first step in algorithm design. The emotion recognition process requires the collection of a large amount of speech data, which should include speech samples in various emotional states. To ensure accurate emotion recognition, the dataset must undergo preprocessing, including denoising and feature extraction. Extracting speech emotion features is crucial for emotion recognition. Commonly used features include intonation, pitch, speech rate, and loudness, which effectively reflect emotional fluctuations.

Common approaches for designing emotion recognition models include traditional machine learning models (Such as support vector machines and decision trees) and deep learning models (Such as convolutional neural networks and long short-term memory networks). Deep learning models demonstrate superior performance in emotion recognition, automatically extracting high-level emotional features from raw audio signals and improving recognition accuracy. Model optimization relies on large amounts of labeled data and repeated training and tuning. By continuously adjusting the network structure and hyperparameters, the emotion recognition system can make accurate judgments in complex emotional scenarios.

The emotion generation algorithm generates corresponding speech output based on the identified emotional state. In the process of emotion generation, it is necessary to consider the diversity of emotional expression and the natural fluency of emotions. Speech synthesis technology combined with deep learning models (Such as WaveNet and Tacotron) can adjust speech parameters such as pitch, speed, and tone based on emotional information based on natural language generation, ensuring that the speech output has emotional color and enhancing the emotional atmosphere of interaction.

4.4 Emotional feedback design in voice interaction

In intelligent voice interaction systems, the design of emotional feedback is a key component in enhancing the user experience. The primary function of emotional feedback is to accurately identify the user's emotional state and promptly adjust the system's voice output, making the interaction more natural and emotionally rich. Such as, when the system detects anger, the speech synthesis module can generate softer, soothing voice output, avoiding a cold or robotic tone. When the user expresses joy, the system can respond to the user's emotional needs by raising the pitch and adopting a cheerful tone, thereby enhancing the enjoyment of the interaction.

Emotional feedback design involves more than just speech generation; it also involves integrating emotional state with voice interaction strategies. The system should select appropriate responses based on different emotional states. For positive emotions such as happiness and excitement, the voice system should respond with a warm tone and a fast speech rate. For negative emotions such as sadness and anxiety, the voice system should adopt a calm and soothing tone. The timeliness and accuracy of emotional feedback are directly related to the user's emotional satisfaction. Therefore, when designing the system, it is necessary to consider the real-time and personalization of emotional feedback to ensure that each interaction can closely match the user's emotional changes and improve the overall interaction effect and user satisfaction.

5 EXPERIMENTATION AND EVALUATION

5.1 System implementation and experimental environment

The intelligent voice interaction system based on affective computing in this study was implemented within a specific hardware and software environment. In terms of hardware, the system utilizes a high-performance voice processing unit and computing platform to ensure real-time and accurate data acquisition, processing, and feedback. Specifically, the system's core hardware includes a highly sensitive microphone array for collecting user voice input and a computer server cluster for running algorithms such as affective computing, speech recognition, natural language processing, and speech synthesis. Furthermore, to enhance the accuracy of emotion recognition, a variety of sensors are used to assist in collecting users' facial expressions or physiological signals. With the support of multimodal input, affective computing is even more precise.

In terms of software, the system is implemented using advanced deep learning frameworks such as TensorFlow and PyTorch, which provide powerful support for emotion recognition and speech generation models. Emotion recognition utilizes a model that combines a CNN with a long short-term memory network (LSTM). This model is capable of processing time series data and extracting emotional features from speech signals. Speech synthesis utilizes a generative model based on Tacotron2 and WaveNet, resulting in natural and fluent speech output and the ability to adjust the emotional tone of speech based on the recognized emotional state. During the experiment, a stable network environment and database management system were configured to ensure efficient data access and processing.

The experimental design and testing methods primarily encompass two aspects: first, system-wide performance testing, focusing on evaluating emotion recognition accuracy, speech synthesis naturalness, and voice interaction response time; and second, interactive testing based on real-world user scenarios, aiming to evaluate the system's emotional computing effectiveness and user experience through actual user feedback. The testing methodology employed a standard emotional speech dataset and actual user voice input data for model training and validation. Furthermore, controlled experiments and A/B testing were used to comprehensively evaluate the system's performance in various scenarios.

5.2 System performance evaluation

Key metrics for evaluating system performance include emotion recognition accuracy and the quality of voice interaction. The evaluation of emotion recognition accuracy primarily relies on comparative analysis of model test results. In terms of emotion recognition accuracy, compared with multiple public emotion datasets, the system achieved over 90% accuracy on a standard test set, significantly exceeding traditional emotion recognition systems. In particular, the system's fine-grained classification capabilities demonstrate a clear advantage in multi-emotion classification (Such as anger, sadness, and joy).

The evaluation of voice interaction effectiveness focuses on the system's performance in real-world applications, primarily examining the naturalness of speech and its ability to express emotions. In the speech synthesis evaluation, user feedback indicates that the system-generated speech sounds more natural than traditional speech synthesis systems and can appropriately adjust for emotional variations, allowing users to clearly perceive the emotional tone of the speech. Through a combination of subjective and objective evaluations, the effectiveness of voice interaction was comprehensively tested. The results demonstrate that the system's emotional expression during interaction significantly enhances user immersion and interactive experience.

User experience surveys and feedback analysis were another key focus of this experiment. Through questionnaires and user interviews, a large amount of user feedback data was collected. Most users reported that the system accurately recognized their emotional state and reflected appropriate emotional tone in voice responses, enhancing the enjoyment and intimacy of their interactions with the system. Furthermore, users generally reported that the inclusion of affective computing effectively alleviated their dissatisfaction during interactions. When faced with unsatisfactory answers, the system's tone became gentler and more understanding, further enhancing their trust and satisfaction.

5.3 Comparative analysis with traditional voice interaction systems

Compared to traditional voice interaction systems, intelligent voice interaction systems based on affective computing demonstrate significant improvements in multiple areas. With the introduction of affective computing, the system can more accurately identify the user's emotional state and adjust voice feedback based on this emotional information, resulting in significant improvements in user experience. Traditional voice interaction systems often respond only to semantically based commands, lacking the ability to recognize and respond to user emotions, making interactions appear stilted and lacking warmth. However, systems based on affective computing, by combining sentiment analysis with speech generation models, can adjust voice parameters such as intonation, speed, and tone based on user emotions, making interactions more natural and fluid. They can even soothe anxious users or elevate their mood through voice tone.

In terms of interaction fluency, affective computing enables the system's voice output to not only respond to emotional changes but also provides appropriate feedback in diverse emotional contexts, avoiding the rigid and repetitive response patterns that can occur with traditional voice systems. User feedback indicates that affective computing makes them feel more caring and understood during conversations with the system, thereby improving the fluidity and naturalness of the interaction.

Furthermore, the introduction of affective computing has significantly improved user satisfaction. In traditional voice systems, users often interact with the system only through mechanical speech, lacking an emotional connection. However, systems based on affective computing can provide users with more personalized and emotional feedback, allowing them to experience a greater sense of human care during their interactions with the system. User satisfaction surveys show that users of voice interaction systems based on affective computing generally have higher overall satisfaction than those using traditional voice systems. This is particularly true when the system responds emotionally, significantly enhancing user emotional resonance.

6 RESULTS AND DISCUSSION

The core achievement of this research is the proposal and implementation of an intelligent voice interaction system based on affective computing. This system not only accurately identifies the user's emotional state but also adjusts voice feedback based on this emotional information, significantly enhancing the user-system interaction experience. Compared to traditional voice interaction systems, this research's innovation lies in the deep integration of affective computing, enabling the system to recognize and respond to users' emotional needs, avoiding the cold and mechanical feedback often associated with previous voice interaction systems. The emotion recognition module, through multi-layered feature extraction and deep learning models, enables the system to accurately identify and classify emotions in a variety of contexts. Furthermore, in terms of speech generation, the affective computing-based speech synthesis model, based on natural language generation, adjusts pitch, speech rate, and tone to make the speech more consistent with the user's emotional state. This system design not only has high practical value but also provides new insights for the future development of voice interaction systems, particularly with significant potential for enhancing user experience and emotional connection.

While this research has achieved some success in integrating affective computing with voice interaction systems, several challenges and limitations remain that need to be addressed. First, accurate emotion recognition remains a challenge, especially in complex emotional scenarios. Current emotion recognition systems primarily rely on surface features of speech signals, such as pitch and speech rate, to determine emotional state. However, these systems can misjudge ambiguous or complex emotional expressions. Such as, when a user is tired or experiencing mixed emotions, the emotional characteristics of their speech may not be as distinct, leading to recognition errors. Furthermore, differences in emotional expression among users of different languages, dialects, or cultural backgrounds can also lead to reduced recognition accuracy, a problem particularly acute in cross-linguistic and cross-cultural applications.

Second, system complexity and real-time performance are also major challenges in current research. Affective computing requires rapid processing and analysis of large amounts of speech data, placing high demands on computing resources and real-time performance. In practical applications, excessive latency in emotion recognition and feedback generation can negatively impact the user experience. This is particularly true in scenarios requiring rapid responses, such as smart homes and virtual assistants, where latency can compromise the smoothness and naturalness of interaction. In addition, although the system's multimodal emotion computing can improve the accuracy of emotion recognition by combining facial expressions, voice, and physiological signals, there are still challenges in hardware costs and real-time data processing.

In the future, voice interaction systems based on affective computing will continue to develop and improve in many areas. First, emotion recognition technology will develop towards higher accuracy and greater robustness. With the continuous advancement of deep learning technology, especially the application of large-scale pre-trained models, affective computing's recognition capabilities will be further enhanced. By combining more multimodal data (Such as facial expressions, eye tracking, voice, and physiological signals), affective computing can more comprehensively identify users' emotional states and provide more accurate emotional feedback to the system. Furthermore, affective computing's contextual awareness capabilities will be further enhanced, enabling the system to understand users' emotional needs in different situations and respond intelligently to changing scenarios.

Second, with the continuous evolution of artificial intelligence technology, voice interaction systems based on affective computing will further improve real-time performance and personalization. Combined with advanced edge computing and distributed computing technologies, future voice interaction systems will be able to complete emotion analysis and voice feedback in a shorter time, reducing system latency and improving interaction fluency. Furthermore, the system's level of personalization will be further enhanced. By learning and analyzing historical user data using deep learning technology, voice interaction systems can provide more customized services based on users' personality traits, emotional preferences, and other information, further improving user satisfaction and user experience.

In addition, the integration of affective computing with other AI technologies will open up new horizons for the development of voice interaction systems. With continued breakthroughs in natural language processing, image recognition, robotics, and other fields, future voice interaction systems will not only be able to understand and respond to user emotions but also possess richer emotional intelligence. Such as, by combining affective computing with emotional reasoning, systems can analyze user emotional trends, predict potential needs, and proactively respond, providing more personalized services. Furthermore, combined with technologies such as virtual reality (VR) and augmented reality (AR), affective computing will enable emotional interactions in more immersive environments, making voice interaction systems more interactive and realistic.

In summary, while intelligent voice interaction systems based on affective computing face certain challenges at the current technological level, they hold enormous potential for future development with the continuous advancement and improvement of related technologies. With the further integration of affective computing and deep learning technologies, voice interaction systems will be able to more accurately perceive user emotions, enhance the naturalness and intelligence of interactions, and truly achieve a more efficient and humanized interactive experience between humans and machines.

7 CONCLUSION

This paper focuses on "Research on the Design of an Intelligent Voice Interaction System Based on Affective Computing" and proposes and implements a voice interaction system that incorporates affective computing technology. Through the introduction of affective computing, the system can accurately identify the user's emotional state and dynamically adjust voice feedback based on the recognition results, making the interaction more natural and humane. The core findings of this study show that affective computing can not only significantly improve the quality of speech recognition and speech synthesis but also enhance the emotional connection between users and the system, improving the interactive experience. In experiments, the system demonstrated good results in both the accuracy of emotion recognition and the emotional expression of speech generation, effectively improving user satisfaction and engagement. In addition, through comparative analysis with traditional voice interaction systems, this study further verified the advantages of affective computing in voice interaction, especially in improving interaction fluency and reducing user emotional barriers.

The results of this research provide a new perspective and approach for the design of intelligent voice interaction systems, which have important practical significance. In the field of intelligent voice interaction, traditional systems primarily rely on voice command recognition and lack understanding and response to user emotions. However, systems based on affective computing can deeply engage with human-computer interaction from an emotional perspective, making the interaction process more responsive to user needs and emotional fluctuations. This not only enhances the user-system interaction experience but also opens the possibility for smart devices to operate in more complex application scenarios. Such as, in smart homes, systems can adjust home device settings based on the user's emotional state, improving the comfort and psychological well-being of family members. In the customer service field, affective computing can enable virtual customer service representatives to more flexibly respond to user dissatisfaction or anxiety, thereby improving customer satisfaction and service quality.

In the future, voice interaction systems based on affective computing will demonstrate broad application prospects in multiple fields. With the continuous advancement of artificial intelligence technology, the integration of affective computing with speech recognition, natural language processing, machine learning, and other technologies will promote the popularization and application of voice interaction systems in smart homes, virtual assistants, education and training, mental health, smart healthcare, and other fields. Furthermore, as people place increasing emphasis on personalized services and emotional experiences, the potential for affective computing applications in personalized recommendations and user sentiment analysis will be further explored. Therefore, this research not only provides new insights for the current development of voice interaction technology but also lays a solid foundation for the innovation and application of future intelligent voice systems.

Although this research has achieved initial results in integrating affective computing with voice interaction systems, there are still areas for improvement. Future research can first further improve the accuracy of emotion recognition, especially in complex, ambiguous, or multi-emotional scenarios. Current emotion recognition technology primarily relies on the acoustic features of speech signals, but emotion is not limited to speech and can also be expressed through multimodal information such as facial expressions and body language. Therefore, future research can explore multimodal emotion recognition technology, combining multiple data sources such as speech, images, and video to improve the comprehensiveness and accuracy of emotion recognition.

In addition, with the advancement of intelligent hardware technology, the real-time performance of affective computing systems will also become a key research focus. In practical applications, affective computing systems often need to process and analyze large amounts of speech data, which places high demands on the system's real-time performance and response speed. Future research can explore more efficient algorithms and optimization techniques to enhance the real-time processing capabilities of affective computing, especially in low-latency, high-concurrency scenarios, to ensure smooth and natural user interactions.

Furthermore, with the development of technologies such as deep learning and reinforcement learning, the integration of affective computing and voice interaction systems is expected to become even more intelligent. In the future, by introducing more adaptive algorithms, the system can continuously optimize its interaction strategies based on user emotional feedback. Such as, the system can learn the user's emotional preferences and continuously adjust the emotional expression of feedback speech, achieving more personalized service. Furthermore, by combining affective computing and artificial intelligence reasoning technologies, the system can not only recognize the user's immediate emotions but also predict future emotional changes and provide appropriate emotional feedback in advance.

In summary, intelligent voice interaction systems based on affective computing have broad future development potential and application prospects, worthy of in-depth research and exploration at both the technical and application levels. By continuously optimizing technologies in emotion recognition, feedback generation, and system real-time performance, the results of this research will provide a sustained impetus for the emotional development of voice interaction systems, ushering in a new era of more natural, intelligent, and human-friendly human-computer interaction.

REFERENCES

[1] Huang, K. L., Duan, S. F., & Lyu, X. (2021). Affective Voice Interaction and Artificial I ntelligence: A research study on the acoustic features of gender and the emotional states of the PAD model. *Frontiers in Psychology*, *12*, 664925.

[2] Pei, G., Li, H., Lu, Y., Wang, Y., Hua, S., & Li, T. (2024). Affective computing: Recent advances, challenges, and future trends. *Intelligent Computing*, *3*, 0076.

[3] Alnuaim, A. A., Zakariah, M., Shukla, P. K., Alhadlaq, A., Hatamleh, W. A., Tarazi, H., ... & Ratna, R. (2022). Human-computer interaction for recognizing speech emotions usin g multilayer perceptron classifier. *Journal of Healthcare Engineering*, *2022*(1), 6005446.

[4] Gervasi, R., Barravecchia, F., Mastrogiacomo, L., & Franceschini, F. (2023). Applications of affective computing in human-robot interaction: State-of-art and challenges for manufact uring. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Enginee ring Manufacture*, *237*(6-7), 815-832.

[5] Lv, Z., Poiesi, F., Dong, Q., Lloret, J., & Song, H. (2022). Deep learning for intelligent human–computer interaction. *Applied Sciences*, *12*(22), 11457.

[6] Alnuaim, A. A., Zakariah, M., Alhadlaq, A., Shashidhar, C., Hatamleh, W. A., Tarazi, H., ... & Ratna, R. (2022). Human-computer interaction with detection of speaker emotions u sing convolution neural networks. *Computational Intelligence and Neuroscience*, *2022*(1), 7 463091.

[7] Li, W., Wu, L., Wang, C., Xue, J., Hu, W., Li, S., ... & Cao, D. (2022). Intelligent coc kpit for intelligent vehicle in metaverse: A case study of empathetic auditory regulation of human emotion. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *53*(4), 2 173-2187.

[8] Alsabhan, W. (2023). Human–computer interaction with a real-time speech emotion recogn ition with ensembling techniques 1D convolution neural network and attention. *Sensors*, *23* (3), 1386.

[9] Šumak, B., Brdnik, S., & Pušnik, M. (2021). Sensors and artificial intelligence methods a nd algorithms for human–computer intelligent interaction: A systematic mapping study. *Sen sors*, *22*(1), 20.

[10] Mejbri, N., Essalmi, F., Jemni, M., & Alyoubi, B. A. (2022). Trends in the use of affecti ve computing in e-learning environments. *Education and Information Technologies*, *27*(3), 3867-3889.

[11] Kumar, S., Haq, M. A., Jain, A., Jason, C. A., Moparthi, N. R., Mittal, N., & Alzamil, Z. S. (2023). Multilayer Neural Network Based Speech Emotion Recognition for Smart As sistance. *Computers, Materials & Continua*, *75*(1).