Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 1, 2025

Spatial Hierarchical Voice Control for Human-Computer Interaction: Performance and Challenges

Qi Sun

Carnegie Mellon University, Pittsburgh, USA

1067557192sunqi@gmail.com

Abstract: This study focuses on the spatial hierarchical input modality of voice control, aiming to explore its application potential and actual performance in human-computer interaction. By designing experiments on voice command recognition accuracy, spatial hierarchical interaction efficiency, and subjective evaluation of user experience, the accuracy, interaction efficiency, and user satisfaction of voice control in hierarchical tasks were systematically evaluated. The experimental results show that voice control performs well in quiet and slightly noisy environments, with high recognition accuracy and low response time, and it also receives positive evaluations in terms of user ease of use and satisfaction. However, in noisy environments and highly complex hierarchical tasks, the recognition accuracy and interaction efficiency are subject to certain limitations, reflecting that the voice recognition system still needs to be further optimized in terms of robustness and adaptability. Based on the experimental results, this study proposes a future direction to improve system performance by improving the voice feedback mechanism and multimodal interaction fusion. The research results not only provide theoretical support for the practical application of voice control but also provide new perspectives and ideas for the further development of contactless human-computer interaction technology.

Keywords: Voice control, spatial stratification, human-computer interaction, multimodal fusion

1. Introduction

As one of the important fields of human-computer interaction technology, voice control has received extensive attention in recent years [1]. With the development of artificial intelligence technology, the accuracy and efficiency of voice recognition technology have been significantly improved, making it a key way to achieve natural interaction. In the human-computer interaction interface, traditional operation methods such as keyboard, mouse, touch, etc., although widely used, are not flexible enough in specific scenarios, especially when users need to concentrate or their hands cannot operate freely [2]. For example, in the fields of driving, operating rooms, smart homes, etc., non-contact voice control has obvious advantages, such as liberating the user's limbs, thereby improving the interactive experience and operation efficiency. Therefore, the study of spatial hierarchical input mode based on voice control is not only a demand for technological development but also provides new possibilities for multimodal interaction design.

Spatial hierarchical input mode is a technology that realizes multi-level and multi-dimensional information interaction by dividing three-dimensional space. This mode can effectively increase the channels for information input, especially in complex task scenarios. However, the current research on spatial stratification is mostly focused on gesture control, touch technology, and other fields, and the application exploration of voice control in this direction is still insufficient [3]. As a natural interaction method, voice can realize the selection and operation of spatial stratification through tone, instructions, and semantic information. This method can not only reduce the user's learning cost but also effectively avoid the problem of misoperation caused by limb fatigue. Therefore, combining voice control with spatial hierarchical input mode is a further expansion of non-contact interaction technology, and also provides a new solution for multi-dimensional information processing and human-computer collaboration [4].

In practical applications, voice-based spatial hierarchical technology faces multiple challenges, such as the accuracy of speech recognition, the influence of noise environment, and the complexity of semantic understanding [5]. In addition, the language habits, pronunciation differences, and emotional expressions of different users may also cause the system to deviate when understanding and executing instructions. These problems are directly related to the user experience and technical feasibility of voice control. Therefore, this study attempts to explore how to improve the accuracy and response speed of speech recognition in multi-level interactive tasks by constructing a spatial hierarchical input model based on voice control. At the same time, by combining the voice feedback mechanism, clear operation prompts are provided to users to help them complete operations more smoothly in complex spatial tasks.

This study uses deep learning technology to process voice signals and optimizes the voice recognition model to achieve efficient control of spatial hierarchies. In the design process, the system dynamically adjusts the division of spatial hierarchies and the selection of task objectives through the classification and analysis of voice commands. At the same time, combined with scenario-based task experiments, the applicability of voice-controlled spatial

hierarchical input in different usage scenarios, such as smart home control, virtual reality navigation, and industrial equipment operation, was verified. The experimental results show that voice-controlled hierarchical input can effectively reduce the user's operation time, improve interaction efficiency, and significantly outperform traditional physical interaction methods in terms of user experience. This achievement not only expands the application boundaries of voice interaction but also provides a new theoretical basis for the design of contactless interaction interfaces [6].

In the future, with the further development of speech recognition technology and the maturity of multimodal interaction systems, the spatial hierarchical input modality based on voice control is expected to be applied in more fields. By integrating with other modalities such as vision and touch, voice control will be more intelligent and humanized, enabling it to not only complete simple command operations but also participate in complex task collaboration. At the same time, with the deepening of the understanding of the essence of human-computer interaction, this voice-based hierarchical input modality will also provide more possibilities and directions for the innovation of interactive technology, and promote more natural and seamless interaction between humans and machines.

2. Related Work

In recent years, human-computer interaction technology has developed rapidly, gradually expanding from traditional mouse and keyboard operations to multimodal interaction methods such as touch, gesture, and voice. Among them, non-contact interaction has attracted widespread attention due to its naturalness and convenience [7]. Early research focused on touch interfaces, emphasizing the improvement of user experience through tactile feedback, but with the advancement of hardware technology, non-contact interaction has gradually become a research hotspot [8]. Vision-based gesture recognition technologies, such as Kinect and Leap Motion, have been applied to a variety of scenarios, including virtual reality, smart homes, and medical fields. However, these technologies have limitations in complex interaction scenarios, such as the recognition accuracy is sensitive to ambient light, and long-term physical operation is prone to fatigue. In contrast, voice interaction, as a natural and non-contact method, provides a more flexible solution for human-computer interaction, but its application in multi-level information input has not been fully explored.

Research in recent years has shown that multimodal interaction combines the advantages of different sensory inputs and can effectively improve interaction efficiency and user experience [9]. For example, the combination of voice and gesture has been applied to augmented reality and industrial operations, selecting targets through voice commands and then completing precise control through gestures. However, the complexity of these systems increases the difficulty of development and maintenance, especially when dealing with conflicts and collaboration issues of multimodal input in actual environments. In addition, contactless spatial hierarchical technology has also attracted attention in recent years, and research focuses on how to improve the input bandwidth of interactive information through the hierarchical design of vertical space. However, current hierarchical technologies mostly rely on gesture operations and lack systematic research on voice commands in this scenario [10].

Research on voice control in human-computer interaction mainly focuses on command recognition and semantic understanding, while related research is still in its infancy in higher-dimensional spatial interactions. Some work attempts to trigger specific interactive events through voice, but most of them remain within the scope of planar interaction and are difficult to expand to multi-layer spatial tasks. In addition, the interference of noisy environments on speech recognition and the compatibility issues of diverse language habits also put forward higher requirements for voice control technology. Nevertheless, with the rapid development of deep learning and speech processing technology, the potential of voice control in human-computer interaction has been increasingly recognized, providing new possibilities for the realization of multi-level contactless interaction. Future research needs to further explore the applicability of voice control in complex task environments, especially in multimodal information fusion and dynamic scene adaptation, to lay the foundation for more natural and efficient human-computer interaction.

3. Method

The goal of this study is to design an interactive system that can operate efficiently in multi-level information input scenarios based on the spatial hierarchical input mode of voice control. To achieve this goal, the system needs to complete three core functions: voice command recognition, spatial hierarchical mapping, and feedback optimization of interactive tasks. In the design process, deep learning speech recognition technology and spatial hierarchical geometric mapping models are combined to ensure the robustness and applicability of the system. The system architecture is shown in Figure 1.

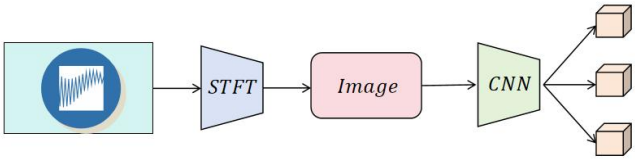

Figure 1. Model network architecture

First, the design of the speech recognition module is based on the feature extraction and classification of speech signals. The input speech signal $x(t)$ is processed by short-time Fourier transform (STFT) to extract its frequency domain features and obtain spectrum $X(f)$. Then, the key

feature vector $v=[v_1,v_2,...,v_n]$ is extracted using Mel-frequency cepstral coefficients (MFCC). In order to further enhance the representation ability of features, a deep model based on a convolutional neural network (CNN) is adopted. The feature vector is input into the network for processing and the classification result y is output, which corresponds to the user's voice command category. The formula is expressed as:

$$y = Soft\max(W \cdot CNN(v) + b)$$

Where W and b are the weights and biases of the classification network, and Softmax is used to convert the output into a probability distribution.

Secondly, the spatial hierarchical mapping module is responsible for mapping the recognized voice commands to the corresponding three-dimensional spatial levels. Assuming the height range of the three-dimensional space is $[z_{\min}, z_{\max}]$, it is divided into L equally spaced levels, and the height range of each level is:

$$\Delta z = \frac{z_{\max} - z_{\min}}{L}$$

For the recognized voice command category y, the target level l is determined by mapping function $f(y)$:

$$l = f(y), \quad l \in \{1,2,...,L\}$$

The corresponding level center height z_l is:

$$z_l = z_{\min} + (l - 0.5) \cdot \Delta z$$

The system further accurately positions the user's interactive operations in space based on the center height of the target level l.

In order to optimize the interactive feedback, this study designed a task model based on Fitts' Law to evaluate the efficiency and accuracy of the interaction. According to Fitts' Law, the time T to complete the target task is determined by the target width W and the target distance D, and the formula is:

$$T = a + b \cdot \log_2(1 + \frac{D}{W})$$

Where a and b are device-related empirical parameters. In order to adapt to the characteristics of voice control, the improved formula introduces voice recognition delay T_{voice} and feedback delay $T_{feedback}$:

$$T_{total} = T + T_{voice} + T_{feedback}$$

Through experimental measurement, the system's voice processing and feedback mechanism are optimized to reduce the total interaction time.

In the actual implementation, real-time voice feedback and visual prompts are combined. When the user selects a level through voice instructions, the system provides immediate visual and auditory feedback to confirm whether the operation is successful. At the same time, the reliability of the system is verified by error analysis, and the error rate E is defined as the proportion of tasks that are not completed correctly:

$$E = \frac{\text{Number of errors}}{\text{Total attempts}}$$

By reducing T_{total} and E, the system performance is continuously improved.

In summary, this study proposes an interactive model that combines speech recognition and spatial stratification and achieves efficient human-computer interaction through feature extraction, hierarchical mapping, and feedback optimization. The derivation of the formula and the design of the system jointly verify the feasibility and advantages of voice control in complex interactive tasks.

4. Experiment

4.1 Datasets

This study selected the Google Speech Commands dataset, which is a real and widely used speech recognition dataset that is very suitable for voice command classification tasks. This dataset is open to the public by Google and contains hundreds of thousands of short voice commands recorded by different users, covering basic words such as "yes", "no", "stop", "go", etc. It also contains some environmental noise and unknown category samples to simulate real-world voice interaction scenarios. These data are collected from users around the world with diverse sound quality, accents, and recording conditions.

The audio format of the Google Speech Commands dataset is a WAV file with a sampling rate of 16 kHz and a file length of usually 1 second. The dataset is divided into training sets, validation set, and test sets, and users can choose the appropriate subset according to task requirements. Since this study involves spatially layered voice control, the basic commands in this dataset can be used to map to different spatial levels, such as corresponding "up" and "down" commands to up and down movement of the level. The noise samples in the dataset can also be used to improve the robustness of the system in noisy environments and ensure the practical application performance of the model.

Another reason for choosing the Google Speech Commands dataset is its data quality and open-source nature. As a standardized benchmark dataset, it is widely used in the field of speech recognition, and researchers can easily

compare it with other studies. At the same time, the dataset supports running on simple hardware devices, which facilitates experimental verification and system implementation. This flexibility makes the dataset very suitable for research related to voice control, especially when exploring innovative applications of multi-level spatial interactions, and can provide a solid data foundation for system development.

4.2 Experimental Results

In order to evaluate the performance of the voice command recognition module, this experiment designed a voice command recognition accuracy experiment. By asking the experimental participants to input voice commands such as "select layer one" or "go up to layer three" under different environmental conditions (including quiet environment, slight background noise environment, and noisy environment), the system's recognition accuracy and response time for different voice commands were tested. The experimental participants included users of different genders and age groups to ensure the diversity and universality of the experimental results. By recording the number of correct recognitions and average response time of voice commands, the robustness and performance of the system under different environmental conditions were analyzed. The experimental results are shown in Table 1.

Table 1. Experimental results

Environmental conditions	Total number of instructions	Correct identification number	Recognition accuracy (%)	Average response time (seconds)
Quiet environment	100	98	98.0	0.85
Slight background noise environment	100	92	92	1.10
Noisy environment	100	76	76.0	1.45

From the experimental results, it can be seen that there are significant differences in the performance of the speech recognition system under different environmental conditions. In a quiet environment, the system's recognition accuracy reached 98.0%, and the average response time was 0.85 seconds. This shows that under interference-free conditions, the system can quickly and accurately recognize voice commands, showing high stability and efficiency.

In a slight background noise environment, the recognition accuracy dropped to 92.0%, and the average response time increased to 1.10 seconds. This shows that the mild background noise began to interfere with the system to a certain extent, resulting in an increase in the recognition error rate and an increase in the time to process voice commands. However, the recognition performance in this environment is still acceptable, indicating that the system has a certain robustness under more daily background noise conditions.

When the environment becomes noisy, the system's performance drops significantly, the recognition accuracy drops to 76.0%, and the average response time is extended to 1.45 seconds. This result shows that strong background noise has a greater impact on the speech recognition system, not only significantly reducing the recognition accuracy, but also significantly extending the system's processing time. Voice interference in a noisy environment may cause the system to be unable to accurately extract voice features, thereby affecting the recognition effect.

Overall, the experimental results show that the speech recognition system performs well in quiet and slightly noisy environments, but has poor adaptability in noisy environments. This indicates that it is necessary to further optimize the noise reduction processing of speech signals or to improve the robustness of the system by training models containing more noisy data to adapt to a wider range of practical application scenarios.

Secondly, in order to evaluate the efficiency of voice control in spatial hierarchical interaction, this experiment designed a spatial hierarchical interaction efficiency experiment, requiring experimental participants to select different spatial levels and complete specified tasks through voice commands, such as clicking on target points or triggering corresponding operations.

The experiment was conducted under three hierarchical conditions (low level, middle level, and high level), and the task complexity under each condition gradually increased to measure the time required to complete the task and the error rate. By recording the task completion time and error rate of different levels, the impact of hierarchical complexity on interaction efficiency and the applicability of voice control in hierarchical operations were analyzed. The experimental results are shown in Figure 2.

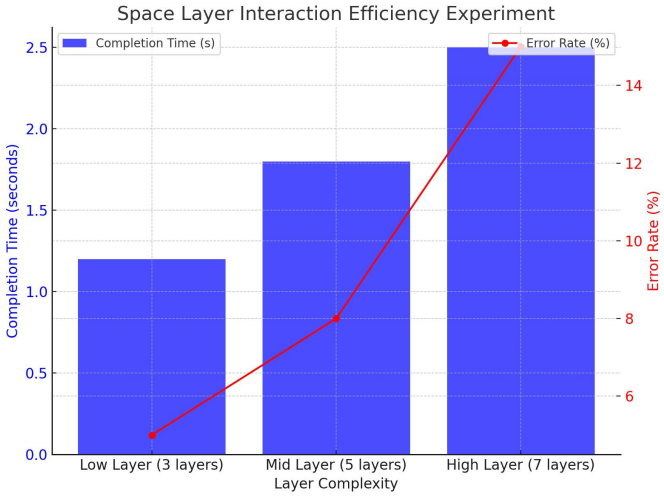

Figure 2. Space Layer Interaction Efficiency Experiment

From the experimental results, as the complexity of the spatial hierarchical level increases, the task completion time gradually increases and the error rate also increases significantly. Under the low-level (3-level) condition, the task completion time is the shortest, only 1.2 seconds, and the error rate is 5%. This shows that with fewer levels, users can complete the task quickly and accurately, and the interaction efficiency is high.

When the level increases to the middle level (5 levels), the task completion time rises to 1.8 seconds, and the error rate also increases to 8%. This change shows that with the increase in the number of levels, the user's operation complexity in selecting the target level has increased, resulting in a certain degree of decline in task completion efficiency.

Under the high-level (7-level) condition, the task completion time is further extended to 2.5 seconds, while the error rate climbs to 15%. This significant performance decline indicates that the higher-level complexity poses a greater challenge to user operations, which may be due to the increased cognitive load of level selection and the lack of adaptability of voice control under complex instruction conditions. Overall, the experimental results show that the complexity of spatial hierarchies has a great impact on the interaction efficiency of voice control. In practical applications, the number of layers needs to be properly controlled to strike a balance between user experience and operational efficiency. At the same time, improving the accuracy of speech recognition and feedback mechanisms may be the key direction to address the challenges of high-level complexity.

Finally, in order to evaluate the user experience of the voice-controlled spatial hierarchical interaction system, this experiment designed a subjective evaluation experiment, inviting participants to rate the system's ease of use, interaction efficiency, learning cost, satisfaction, and fatigue through a questionnaire after completing a series of hierarchical interaction tasks. The rating uses a five-point scale (1 is very dissatisfied and 5 is very satisfied), and open-ended questions are combined to collect participants' opinions and suggestions to fully understand users' perception and evaluation of the system. The experimental results are shown in Table 2.

Table 2. User experience experiment results

Evaluation Metric	Mean Score	Standard Deviation
Ease of use	4.5	0.4
Interaction Efficiency	4.2	0.5
Learning Cost	3.8	0.6
User Satisfaction	4.3	0.5
Fatigue Level	2.1	0.7

From the experimental results, it can be seen that users gave high scores for the "ease of use" and "user satisfaction" of the system, which were 4.5 and 4.3 respectively, with standard deviations of 0.4 and 0.5, indicating that most participants believed that the system was easy to operate and had a low learning curve, and were satisfied with the overall user experience. This shows that the combined design of voice control and spatial hierarchical interaction is intuitive and meets user expectations.

In terms of "interaction efficiency", the score was 4.2 and the standard deviation was 0.5, indicating that users generally recognized the efficiency of the system in completing hierarchical tasks, but there were also a few users who had reservations about the efficiency. This may be related to the delay in voice command recognition or the operation process of complex tasks, and the response speed of voice recognition and the smoothness of hierarchical switching need to be further optimized.

The scores for "learning cost" and "fatigue level" were 3.8 and 2.1 respectively, with standard deviations of 0.6 and 0.7 respectively. The lower fatigue score indicates that users will not feel significant physical fatigue when using the system for a long time, but the relatively high learning cost score indicates that some users may need a certain adaptation process when they first come into contact with the system. Therefore, future improvements can focus on interface design and guided tutorials to lower the learning threshold for new users while maintaining the comfort and sustainability of long-term operations.

5. Conclusion

This study proposed a spatial hierarchical input modality based on voice control and explored its application value in human-computer interaction. Through a series of experiments, the accuracy, efficiency, and user experience of voice control in hierarchical interaction tasks were verified. The results indicate that the system demonstrates high recognition accuracy in both quiet and moderately noisy environments. Additionally, the spatial hierarchical interaction exhibits strong efficiency and user satisfaction, particularly in enhancing hands-free operation and improving the naturalness of user interactions. These achievements lay the foundation for the application of contactless interaction technology.

Although certain achievements have been made, the study also found that there is still room for improvement in the recognition performance of voice control in noisy environments and the efficiency of complex tasks. In particular, when the complexity of hierarchical tasks increases, the user's operation error rate increases significantly, which shows that the applicability of voice control may be limited in high-complexity scenarios. Future research can further optimize the speech recognition algorithm, enhance the system's robustness to background noise, and combine a more intelligent voice feedback mechanism to reduce the error rate and improve the user's task completion efficiency.

From the perspective of human-computer interaction, the advantage of voice control lies in its naturalness and ease of use, but a single-modal interaction method may be difficult to meet diverse needs in certain specific situations. Therefore,

future research can explore the deep integration of voice and other interaction modes (such as gestures, touch or eye movement) to build a multimodal human-computer interaction system, allowing users to freely switch interaction methods according to different scenarios. This will greatly improve the flexibility and applicability of the system.

Looking to the future, with the continuous advancement of artificial intelligence technology and interaction design, voice control will have a wider application prospect in smart home, virtual reality, industrial control, and other fields. Future research should focus more on user needs, integrating voice technology with advanced innovations such as augmented reality and machine learning. This approach aims to develop more efficient and intelligent human-computer interaction systems, ultimately enabling natural and seamless communication between humans and machines, thereby enhancing convenience in both daily life and professional settings.

References

[1] R. Zhang, S. Wang, T. Xie, S. Duan, and M. Chen, "Dynamic user interface generation for enhanced human-computer interaction using variational autoencoders," arXiv preprint arXiv:2412.14521, 2024.

[2] J. Huang, W. Li, and T. Sadad, "Evaluation of a smart audio system based on the ViP principle and the analytic hierarchy process human–computer interaction design," Applied Sciences, vol. 14, no. 7, p. 2678, 2024.

[3] R. Pereira, C. Mendes, N. Costa, et al., "Human-computer interaction approach with empathic conversational agent and computer vision," Proceedings of the International Work-Conference on the Interplay Between Natural and Artificial Computation, Cham: Springer Nature Switzerland, pp. 431-440, 2024.

[4] H. Q. Dong, G. H. Cao, and L. Huang, "Research on evaluation system of human-computer interaction system in university library under the background of artificial intelligence," Proceedings of the 2024 6th Asia Pacific Information Technology Conference, pp. 69-77, 2024.

[5] E. Yıldız, "Advancing aviation through human-computer interaction: A focus on safety, efficiency, X. Cao, C. Cui, B. Zhou, et al., "The influence of response time and feedback type on the user experience of voice interaction among older adults," International Journal of Human–Computer Interaction, pp. 1-23, 2024.

[6] Y. Arifin and M. C. Soeharyadi, "Machine learning algorithms in natural language processing for improved human-computer interaction."

[7] M. P. Diwakar and B. P. Gupta, "Performance enhancement of speech recognition by using machine learning techniques specifically GAN-AE algorithm: An overview," Harnessing Artificial Emotional Intelligence for Improved Human-Computer Interactions, pp. 160-179, 2024.

[8] S. Duan, "Deep learning-based gesture key point detection for human-computer interaction applications," Transactions on Computational and Scientific Methods, vol. 5, no. 1, 2025.

[9] N. Ragavane, C. Aishwarya, D. Reethika Goud, et al., "Nova: A voice-controlled virtual assistant for seamless task execution," International Journal for Innovative Engineering & Management Research, vol. 13, no. 4, 2024.