Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 2298–2308
https://doi.org/10.54941/ahfe1005798

Exploring the Impact of Error Feedback Methods on User Experience in Voice Interaction

Tongtong Xie[1], Meng Li[1], Yinyin Bai[2], Jieren Xie[1], Aibin Zhu[1], and Zengyao Yang[1]

[1]School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China
[2]School of Medicine and Forensics, Xi'an Jiaotong University, Xi'an, China

ABSTRACT

Voice interaction has played an important role in various scenarios such as smart homes, cars, and healthcare due to its ease of use, efficiency, and convenience. However, errors in voice interaction can greatly affect the user experience. This paper aims to explore the impact of different error feedback methods on user experience, with the goal of improving the user experience of voice interaction. The study utilizes a combination of subjective and objective approaches by creating an experimental platform to collect facial expression data and emotional valence evaluation data from participants. By analyzing the data, user preferences for different feedback methods can be determined. The findings suggest that in directive task scenarios, users prefer feedback that directly explains the error. In broadcast task scenarios, users prefer feedback that explains the error and provides a commitment to resolve it. In conversational task scenarios, users prefer intelligent voice assistants to take the lead in the conversation, guide its direction, and provide specific event suggestions. This research contributes to a better understanding of the impact of error feedback methods on user experience and provides guidance and reference for the design of future error feedback methods in voice interaction.

Keywords: Voice user interaction, Error feedback, Emotion recognition, Emotion value

INTRODUCTION

The interaction between humans, machines, and objects is becoming an essential aspect of the information society, having a profound influence on human production and daily life. Among the various forms of interaction, voice interaction has played a crucial role in home automation, in-car systems, and healthcare due to its user-friendly, efficient, and convenient nature. However, voice interactions often suffer from machine errors and misunderstandings, resulting in a compromised user experience. An analysis of the impact of intent and recognition errors on the interaction experience reveals that effective feedback methods can alleviate negative user experiences. Therefore, while continually optimizing voice interaction algorithms to minimize errors, exploring the influence of error feedback methods on user experience can enhance error handling and foster user trust, ultimately increasing user engagement.

Scholars have conducted research on error feedback methods before. Current error feedback patterns in voice interaction primarily consist of two elements: apology and responsibility allocation. The experiments showed that sincerely acknowledging errors can alleviate negative user emotions (Mahmood et al., 2022; Dabre et al., 2020). And concise and straightforward feedback after errors is more preferred by users (Hass et al., 2022). Additionally, user preferences and expectations regarding feedback methods vary depending on the task at hand. When completing factual and explanatory tasks, users prefer the product to explain the cause of the error and express apologies. However, when performing complex exploratory tasks, users also expect the product to provide operational guidance (Yuan et al., 2020). Moreover, users often rely on voice interaction for task guidance (Myers et al., 2018). Furthermore, there is a diminishing marginal utility in the feedback process when the same feedback statement is repeated more than three times(Kim et al., 2021), the alleviating effect on negative emotions significantly decreases.

Extensive literature research highlights a gap in studies that comprehensively explore the impact of error feedback on user experience through a combination of subjective and objective approaches. Therefore, the objective of this paper is to investigate the influence of various error feedback methods on user experience in voice interaction, specifically focusing on the usage scenario of smart speakers in home environments. This study will establish a model for error feedback in everyday voice interaction tasks by conducting a survey of prevalent smart speaker models available in the market. Additionally, a user experience testing platform for voice interaction will be developed, utilizing facial expression recognition and rating scale questionnaires for data analysis. The findings of this study will serve as valuable design references for future error feedback methods in voice interaction.

METHODS

Task Classification and Importance Accessment

Different usage scenarios give rise to diverse user needs and expectations for speech interaction, which, in turn, influence the user experience following errors. Researchers such as Chelsea Myers and Anushay Furqan have identified four primary reasons for speech interaction errors: ambiguous intent, natural language processing error, failed feedback, and system error (Myers et al., 2021). System errors are algorithmic errors that are currently irremediable through feedback improvements; therefore, the analysis primarily focuses on the first three causes of errors. Currently, the top-selling smart speakers in China include Xiaodu, Tmall Genie, and Xiaoai. This paper selects voice tasks from these three smart speakers in real-life scenarios as experimental tasks and categorizes them based on user interaction needs. The impact of the same feedback on user experience can vary depending on the type of speech interaction task. For instance, an apology for failing to turn on the lights and resulting in the user's tardiness due to an improperly set alarm clock will have different mitigating effects

on user experience. Thus, this paper will also evaluate the significance of interaction tasks.

Voice interaction tasks can be classified into the following three types based on user interaction needs: 1) Directive tasks: These tasks require the smart speaker to receive explicit instructions from the user to fulfill specific requirements in a particular scenario, such as turning on the lights. In this type of task, user needs are well-defined, and tolerance for errors is minimal. 2) Broadcast tasks: These tasks involve the smart speaker providing knowledge and information broadcasts to the user in response to instructions, such as weather queries. Users can accomplish these tasks through alternative means and can tolerate a certain degree of errors. 3) Conversational tasks: These tasks revolve around the smart speaker providing leisure and chat functions, like casual conversations. As users do not have specific goals for these tasks, the accuracy requirements are the lowest.

To sum up, the importance ranking of the three aforementioned speech interaction task types is: directive tasks > broadcast tasks > conversational tasks. The subsequent experiments will explore the impact of error feedback on the user experience for each of these task types.

Experimental Task Scenarios and Feedback Corpus

Three task scenarios were devised for this exploratory experiment, all gleaned from real-life situations. Each scenario encompassed directive tasks, broadcast tasks, and conversational tasks. The aim was to familiarize participants with the task functions and provide an experience reflective of real-life scenarios.

The experimental tasks and corresponding feedback methods for this experiment can be summarized in Table 1 below (see Table 1).

Table 1. Task and feedback.

Scenario	No.	Type	Instruction	Method	Corpus
Scenario 1	1–1	Directive Task	Turn the light on.	Explain the reason.	Feedback 1-1
	1–2	Broadcast Task	Recommend a TV brand.	Encourage the users to repeat.	Feedback 2-1
	1–3	Conversational Task	Vent your grievances.	Give advice to the users.	Feedback 3-1
Scenario 2	2–1	Broadcast Task	Query navigation.	Response implicitly.	Feedback 2-2
	2–2	Directive Task	Play a song.	Encourage the users to repeat.	Feedback 1-3
	2–3	Conversational Task	Share something fun.	Cheer the users up.	Feedback 3-3
Scenario 3	3–1	Conversational Task	Chat with voice assistant.	Take responsibility and encourage to talk	Feedback 3-2
	3–2	Broadcast Task	Check the weather.	Apologize and promise.	Feedback 2-3
	3–3	Directive Task	Book a flight ticket.	Response humorously.	Feedback 1-2

After determining the tasks and corresponding feedback methods, this study formulated the corresponding feedback corpus for each feedback type as follows:

Feedback 1-1: Sorry, I couldn't find any controllable devices.

Feedback 1-2: Sorry, the answer is missing. Please don't mind.

Feedback 1-3: What did you just say? Could you please try saying it more clearly?

Feedback 2-1: Sorry, my little brain didn't understand. Could you please say it again?

Feedback 2-2: That's a beautiful place. Let's go there together sometime.

Feedback 2-3: Sorry, I don't currently support this feature. I'll let you know as soon as it becomes available.

Feedback 3-1: If you're feeling unhappy, try doing something that brings you joy.

Feedback 3-2: Is there something on your mind that you'd like to share with me? Maybe I can help you solve it.

Feedback 3-3: Just keep going, and there's nothing that can't be overcome.

Experimental Platform

The experiment utilized a self-built voice interaction error feedback platform. The platform comprised three interfaces, with each interface corresponding to a different task type. The language data for various feedback methods was displayed within each interface. Feedback corpus required for the study was pre-synthesized and embedded into the code files using speech synthesis software. The experimenter could provide specific error feedback to the participants by clicking buttons. OpenFace software was integrated with the platform to obtain real-time facial expression data from users during the experiment (see Figure 1).

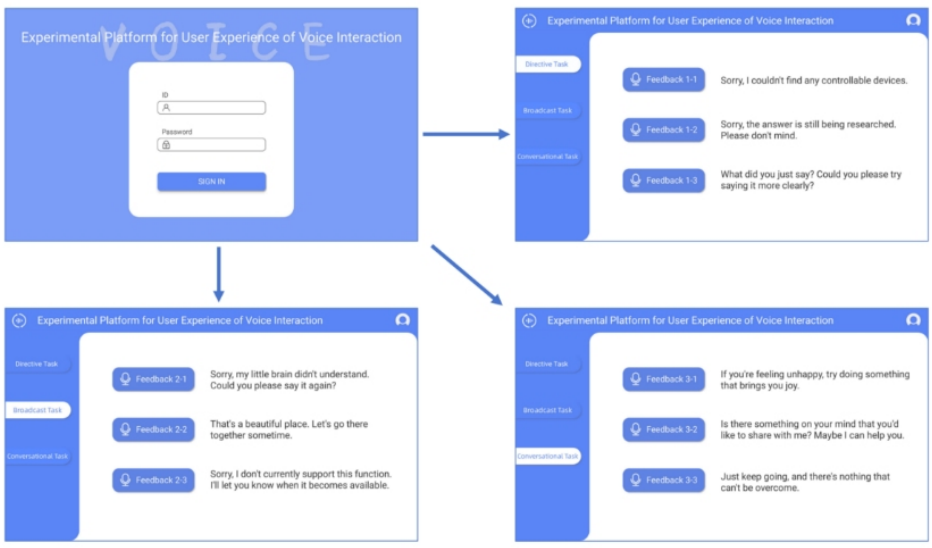


Figure 1: Experimental platform for user experience of voice interaction.

An Asus ZenBook laptop served as the primary equipment for this experiment, equipped with a self-built experimental platform for sending feedback corpus, the OpenFace platform for capturing facial data, and a feedback corpus synthesized from 9 pre-recorded audio files. The experiment took place in a quiet and comfortable indoor environment to prevent noise interference that could affect the experimental process or cause emotional disturbance to the participants.

Experiment

The study recruited 20 participants (10 males and 10 females) who met the age requirements and had no hearing or visual impairments. Additionally, participants had prior experience utilizing voice assistants for task settings.

The experiment comprised three rounds, with each round simulating a continuous speech interaction scenario. Each speech interaction scenario encompassed three tasks, totaling nine tasks per participant. The experimenter assigned the tasks to the participants, and the feedback was delivered through a pre-built platform.

Following the completion of facial data collection, the participants were requested to complete a Likert scale questionnaire. The study utilized the emotion-valence model (see in Figure2), which simplified the PAD model (pleasure-arousal-dominance) initially proposed by Mehrabian and Russell in 1974 (Begany et al., 2015; Zhang et al., 2018). The questionnaire primarily aimed to investigate the participants' feedback evaluations in terms of pleasure, arousal, trust, and novelty.

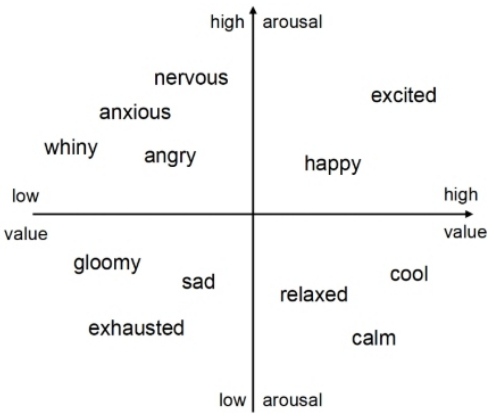


Figure 2: Emotion-valence model.

The experiment procedure is shown in Figure 3(see Figure 3). A pre-experiment was conducted before the formal start of the experiment to familiarize users with the experimental procedure and ensure the smooth progress of the experiment.

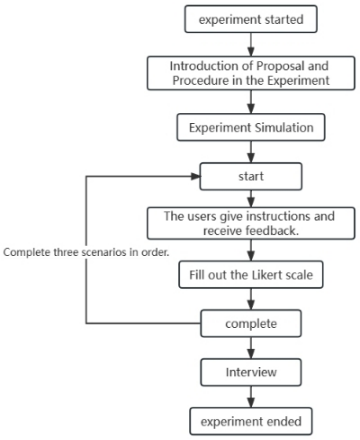


Figure 3: Experiment procedure.

RESULTS

Facial data in this study was collected using the OpenFace software, and data analysis for the Likert scale was performed using IBM SPSS 25.0 software.

Overall Preferences

This study examined users' preferences and the reasons behind them for three distinct feedback methods in directive tasks, broadcast tasks, and conversational tasks, following the completion of the experiment. The findings can be summarized as follows: 1) In directive tasks, 50% of the participants preferred the feedback method of "Explaining the reason." 2) In broadcast tasks, 65% of the participants favored the feedback method of "Apologize and promise." 3) In conversational tasks, 55% of the participants showed a preference for the feedback method of "Taking responsibility and encourage to talk." No significant differences were found in the overall preferences for different feedback methods among participants of different genders in directive tasks ($p = 0.470$), broadcast tasks ($p = 0.666$), and conversational tasks ($p = 0.700$).

Facial data from users was collected using OpenFace, which underwent grayscale conversion, data augmentation, and normalization as part of the image preprocessing process. Emotion classification of the facial data was performed using deep feature learning and recognition techniques (Savin et al., 2022; Moschona et al., 2020). Figure4 shows the procedure of collection the facial data(see Figure 4).

When users experience feelings of comfort and pleasure upon receiving feedback, it can effectively mitigate their negative emotions. And we can see presents the statistical results and proportions of users who experienced comfort or pleasure in response to various feedback types, where the horizontal coordinate is the number of the feedback method, and the vertical coordinate is the number of users who feel happy or comfort(see Figure 5).

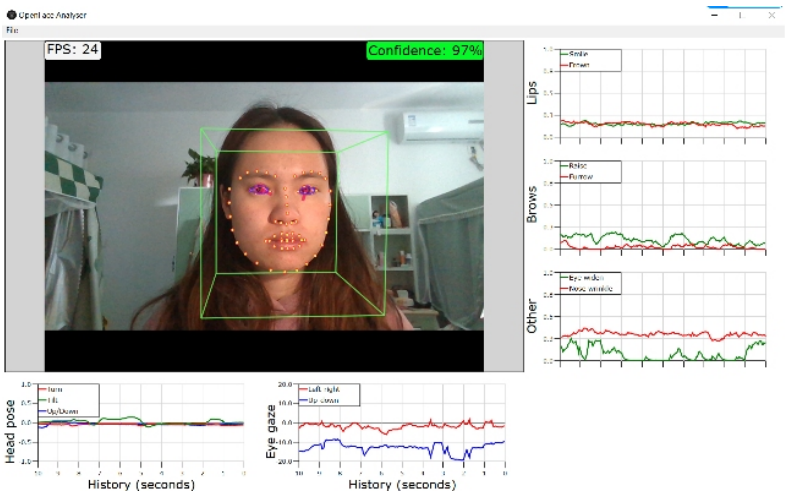
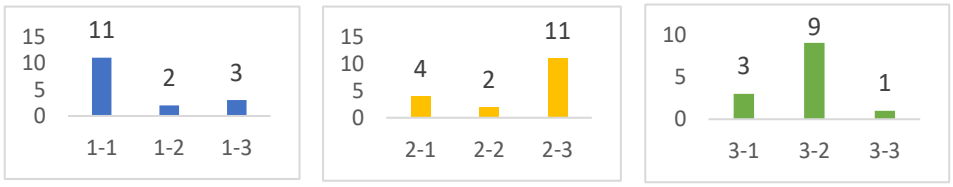


Figure 4: Procedure of collecting the facial data.



(a)Directive Task (b)Broadcast Task (c)Conversational Task

Figure 5: The number of users who felt happy or comfort.

Directive Task

Non-parametric tests were employed to analyze the experimental data, and the findings indicate no statistically significant differences among the four dimensions of the three distinct feedback methods implemented in command-based tasks. However, significant distinctions emerged among the various dimensions of the feedback methods being investigated. Specifically, there were significant differences observed in relation to explaining the error reasons ($p = 0.000$), encouraging users to resend commands ($p = 0.007$), and humorous responses ($p = 0.003$). The results of pairwise comparisons are provided in Table 3-2. It is worth noting that under the "explaining the error reasons" feedback method, there were significant differences between novelty and arousal ($p = 0.001$), as well as trust and arousal ($p = 0.029$). Similarly, under the "encouraging users to resend commands" feedback approach, a significant difference was observed between novelty and arousal ($p = 0.024$). Lastly, employing the "humorous responses" feedback method also demonstrated a notable difference between novelty and arousal ($p = 0.016$); however, no significant differences were identified among the other dimensions. The results of pairwise comparisons in directive tasks are shown in Table II below (see Table 2).

Table 2. The results of comparisons in directive tasks.

Feedback	Explain the Reason		Encourage the Users to Repeat.		Response Humorously.	
	Sig.	adj.Sig.	Sig.	adj.Sig.	Sig.	adj.Sig.
novelty-trust	0.358	1.000	0.391	1.000	0.198	1.000
novelty-pleasure	0.221	1.000	0.245	1.000	0.178	1.000
novelty-arousal	0.000**	0.001**	0.004*	0.024*	0.003*	0.016*
trust-pleasure	0.759	1.000	0.759	1.000	0.951	1.000
trust-arousal	0.005*	0.029*	0.043*	0.260	0.086	0.518
pleasure-arousal	0.012*	0.072	0.086	0.518	0.098	0.589

Based on the qualitative data analysis from post-experiment user interviews, it was revealed that users experienced a negative user experience due to errors in voice interaction. However, various feedback methods were found to have no significant impact in enhancing the user experience. In such instances, users tended to promptly terminate the voice interaction and explore alternative solutions. Users found that explaining the reasons for errors in the feedback helped them resolve issues.

Broadcast Task

Non-parametric tests were employed to analyze the experimental data, which indicated no statistically significant differences in pleasure, arousal, trust, and novelty across the three distinct feedback methods used in broadcast-based tasks. However, Table 3revealed a significant difference in various dimensions within the "encouraging users to resend instructions" feedback method ($p = 0.001$). Contrastingly, significant differences were observed between novelty and arousal ($p = 0.020$) as well as trust and arousal ($p = 0.029$) under the "explaining the error reasons and making promises" feedback method (see Table 3).

Table 3. The results of comparisons in broadcast tasks.

Feedback	Encourage the Users to Repeat.	
	Sig.	adj.Sig.
novelty-trust	0.903	1.000
novelty-pleasure	0.391	1.000
novelty-arousal	0.003*	0.020*
trust-pleasure	0.462	1.000
trust-arousal	0.903	1.000
pleasure-arousal	0.391	1.000

Participants in the post-experiment user interviews conveyed that the smart speaker's provision of error explanations and promises effectively mitigated their feelings of tension and uncertainty towards the interaction outcomes. Conversely, the implementation of novel feedback methods and content engendered a sense of diminished control over the interaction process, ultimately diminishing users' expectations regarding the product's usability.

Conversational Task

Non-parametric tests were performed on the experimental data, revealing no statistically significant differences in pleasure, arousal, trust, and novelty among the three distinct feedback methods employed in open-ended tasks. However, an analysis of the data in Table 4 demonstrates a notable variance in several dimensions for the feedback method of "taking responsibility and encouraging conversation" ($p = 0.001$). Notably, novelty and arousal differed significantly ($p = 0.042$) within the feedback method of "taking responsibility and encouraging conversation"(see Table 4)

Table 4. The results of comparisons in conversational tasks.

Feedback	Take Responsibility and Encourage to Talk	
	Sig.	adj.Sig.
novelty-trust	0.426	1.000
novelty-pleasure	0.098	0.589
novelty-arousal	0.007*	0.042*
trust-pleasure	0.391	1.000
trust-arousal	0.058	0.346
pleasure-arousal	0.298	1.000

In the post-experiment user interviews, participants expressed a preference for the intelligent voice assistant to take the lead in conversational task scenarios. They indicated that the assistant should assume an active role in leading the conversation, guiding its direction based on the user's instructions, and providing specific event suggestions. However, it was noted that a considerable number of users were unfamiliar with this scenario and feedback method. Hence, more research is required to develop effective feedback methods for this type of scenario.

CONCLUSION

This study found no significant gender-based differences in user preferences for feedback methods in various task scenarios. This study also found feedback methods do not significantly enhance the user experience, as they fail to adequately address errors and unmet user needs. Nevertheless, certain feedback methods can alleviate users' negative emotions and increase their willingness to retry the function. For directive task scenarios, users favor feedback methods that provide explanations for error reasons. In broadcast task scenarios, users prefer feedback methods that not only explain the error

reasons but also offer promises. However, innovative feedback methods and content may cause users to perceive a loss of control, resulting in a diminished user experience. In conversational task scenarios, users anticipate the intelligent voice assistant assuming a proactive role by leading the conversation and offering specific event recommendations.

To overcome the limitations of this study, several future research directions are suggested. Firstly, it is recommended to further investigate systematic error, which represent the most fundamental error in voice interaction, to determine their impact on the user experience and their relationship with the other three error types. Secondly, future research should explore alternative task scenarios in public places like schools and hospitals. Lastly, it is recommended to target the elderly and children in future research to develop feedback methods that are suitable for their respective age groups in voice interaction.

ACKNOWLEDGMENT

This work was carried out with the support of Qinchuangyuan Project No. 2024QCY-KXJ-189 Project Name: AI non-contact ocular axial measurement technology.

REFERENCES

Begany, G. M., Sa, N. and Yuan, X. (2015). Factors Affecting User Perception of a Spoken Language vs. Textual Search Interface: A Content Analysis. pp.iwv029–iwv029. doi: https://doi.org/10.1093/iwc/iwv029.

Dabre, P., Gonsalves, R., Chandvaniya, R. and Nimkar, A. V. (2020). A Framework for System Interfacing of Voice User Interface for Personal Computers. 2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA). doi: https://doi.org/10.1109/cscita47329.2020.9137809.

Haas, G., Rietzler, M., Jones, M. and Rukzio, E. (2022). Keep it Short: A Comparison of Voice Assistants' Response Behavior. CHI Conference on Human Factors in Computing Systems. [online] doi: https://doi.org/10.1145/3491102.3517684.

Kim, S., Abhinav Garlapati, Lubin, J., Amir Tamrakar and Ajay Divakaran (2021). Towards Understanding Confusion and Affective States Under Communication Failures in Voice-Based Human-Machine Interaction.arXiv (Cornell University). doi: https://doi.org/10.1109/aciiw52867.2021.9666238.

Mahmood, A., Fung, J. W., Won, I. and Huang, C.-M. (2022). Owning Mistakes Sincerely: Strategies for Mitigating AI Errors. CHI Conference on Human Factors in Computing Systems. doi: https://doi.org/10.1145/3491102.3517565.

Moschona, D. S. (2020). An Affective Service based on Multi-Modal Emotion Recognition, using EEG enabled Emotion Tracking and Speech Emotion Recognition. [Online] IEEE Xplore. doi: https://doi.org/10.1109/ICCE-Asia49877.2020.9277291.

Myers, C., Furqan, A., Nebolsky, J., Caro, K. and Zhu, J. (2018). Patterns for How Users Overcome Obstacles in Voice User Interfaces. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. doi: https://doi.org/10.1145/3173574.3173580.

Myers, C. M., Fernando, L., Acosta-Ruiz, A., Alessandro Canossa and Zhu, J. (2021). 'Try, Try, Try Again:' Sequence Analysis of User Interaction Data with a Voice User Interface. doi: https://doi.org/10.1145/3469595.3469613.

Savin, A. V., Sablina, V. A. and Nikiforov, M. B. (2021). Comparison of Facial Landmark Detection Methods for Micro-Expressions Analysis. doi: https://doi.org/10.1109/meco52532.2021.9460191.

Yuan, S., Brüggemeier, B., Hillmann, S. and Michael, T. (2020). User Preference and Categories for Error Responses in Conversational User Interfaces. Proceedings of the 2nd Conference on Conversational User Interfaces. doi: https://doi.org/10.1145/3405755.3406126.

Zhang, W., Yang, H. and Zhi, P. (2018). Emotional speech synthesis based on DNN and PAD emotional state model. doi: https://doi.org/10.1109/iscslp.2018.8706656.