

RESEARCH ARTICLE

Waiting Experience: Optimization of Feedback Mechanism of Voice User Interfaces Based on Time Perception

JUNFENG WANG¹, YUE LI², SHUYU YANG³, SHIYU DONG¹, AND JIALIN LI¹¹College of Design and Innovation, Shenzhen Technology University, Shenzhen 518118, China²China General Nuclear Power Group, Shenzhen 518038, China³School of Manufacturing Science and Engineering, Southwest University of Science and Technology, Mianyang 621010, China

Corresponding author: Junfeng Wang (wangjunfeng@sztu.edu.cn)

This work was supported in part by the Humanities and Social Science Research Planning Fund of the Ministry of Education of the People's Republic of China through "Research on Adaptive Interaction Design of Intelligent Speech Products" under Grant 21YJC760078.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Committee of Shenzhen Technology University under Application No. YJ202209, and performed in line with the Medical Ethics.


ABSTRACT Waiting is an indispensable and inevitable part of man-machine voice interaction. The voice user interface (VUI) feedback mechanism is a key factor affecting voice interaction's waiting experience. The feedback time of most available voice interfaces is fixed or decided by the processing time of hardware and software, which has not been designed and cannot offer users a good interaction experience. In this paper, the speech rate of user-machine voice interaction is collected through prototype experimentation. Besides, users' time perception of different voice interfaces' feedback time settings is studied based on time psychology theories. Moreover, users' emotional changes are described after a specific feedback time with the distribution of two-dimension arousal-valence emotion space. Users' time perception and subjective emotions are differently influenced by different VUI feedback times. The experimental results show that 750 ms is the optimal VUI feedback time point at which the best users' subjective feelings and psychological experiences are reached, and the threshold limit time spent by users in waiting for the VUI feedback is 1,850 ms which will lead to user emotions with low levels of arousal and valence after being exceeded. Based on that, a linear regression model is proposed to define the optimal feedback time of VUI. The user experience VUI research results show that the calculated feedback time parameters can make users produce time perception in line with their expectations in interacting with voice interfaces.

INDEX TERMS Voice user interface, feedback time, time perception, speech rate.

I. INTRODUCTION

In recent years, research on voice interaction and voice user interface (VUI) was focused on implementing hardware and technology. In contrast, relatively less research was conducted on users' interaction experience brought by VUI. Feedback time is essential to the interaction between VUI voice assistant applications and users. Different feedback time parameters bring users psychological feelings such as

the sense of urgency and delay and significantly impact users' experience, one of the subjects deserving attention and research. Unlike the graphical user interface (GUI), in which one state can be maintained on the screen for any time, time-series association rules generally used in people's dialogue should be followed in VUI. When people talk, dialogue with too short a delay will bring listeners a sense of stress and hurry. At the same time, silence for a long time or response after a long period will distract or confuse the user and then destroy that dialogue. Therefore, time strategies available for people rather than technically feasible strategies should be applied in VUI [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Giuseppe Desolda .

Voice user interface is also called voice interface or auditory interface, which refers to the interface using sound (including speech and no-speech sound) to realize the input, output, feedback, and response of information [2]. This research focuses on voice interfaces in which the user talks to the device, and the device responds with a synthesized voice. Current main carriers of man-machine voice interaction are voice assistant applications based on VUI, such as intelligent voice assistants like Siri of Apple and “Xiaoai” of Xiaomi, and artificial intelligence (AI) assistants like Bixby of Samsung and Cortana of Microsoft.

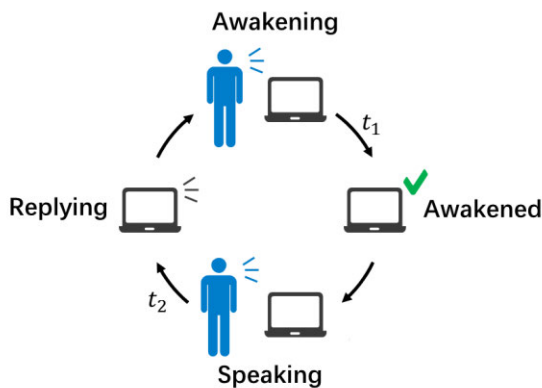


FIGURE 1. The cyclic process of VUI voice interaction.

As shown in Figure 1, the VUI-based voice interaction between user and voice assistant applications can be divided into four phases: 1) users speak specific awakening words to awaken VUI, if the words could be recognized by VUI as a predefined instruction, 2) VUI present the state of being awakened and listening, 3) users give a voice command or say a word to VUI, and 4) VUI determines whether the user has finished a sentence based on the interval. If the VUI believes that the user has finished the instruction, it recognizes what has been said and replies to the user. In the process of voice interaction, the feedback time lies in two phases; the first is the total time, t_1 , between “users speaking specific awakening words to awaken VUI” and “VUI presenting the state of being awakened.” The second is the total time, t_2 , between “users giving the voice command to VUI” and “VUI replying to users.” In human conversation, a speaker may say more than one sentence to express his/her mind or raise a question, while the listener needs to understand the speaker and judge whether the speaker has finished his/her words [3], [4]. Similarly, VUI needs to analyze the voice signals and pauses gaps in human speech to determine if the user has finished a word [5]. Voice activity detection (VAD) is widely used to solve this problem [6], [9].

The actual feedback time of VUI consists of two parts. One is the time spent processing a user’s commands by software/hardware. Another is timely and intuitive feedback close to the reply in human-to-human interaction. The former may present constraints on the latter simply due to

hardware/software processing limitations that can affect timely response. However, users may not notice this two-phases protocol. They just speak and wait for the response from VUI. The waiting time is the principal factor affecting users’ time perception, and an appropriate setting of actual feedback time is a major way to improve users’ time perception, which affects the user experience of VUI. Currently, the feedback time of most voice assistant applications based on VUI available on the market is fixed or decided by hardware performance. So, in many cases, these applications cannot bring users favorable subjective feelings and time perception. With the advances in voice interaction technology, apparent homogeneity and substitutability have been shown among various voice assistant applications based on VUI. Therefore, users’ experience when using these applications becomes particularly important, and the VUI feedback time might decide whether users will stay. The crucial concern for VUI designers should be how to set the feedback time of VUI to bring users positive feelings and a satisfying time experience in voice interaction.

In this paper, based on basic theories of time psychology and prototype experimentation, the effects of different feedback times of VUI on users’ time perception and the dependence relation between users’ time perception and speech rate were studied through VUI voice interaction experiment stimulation. Meanwhile, the “optimal feedback time” definition model was built based on experimental data. Finally, the availability of the “optimal feedback time” model was verified through prototype experiments, and design strategies were proposed based on experimental results for developing existing VUI voice assistant applications.

II. RELATED WORK

This section will give a brief overview of literature related to our research, including emotion in user voice interaction and time perception.

A. EMOTION IN USER VOICE INTERACTION

In the process of user voice interaction, the user’s emotion could be affected by the response of VUI, which obtained some researcher’s interest. Swoboda et al. [10] compared the effectiveness of physiological features and speech features to predict the intensity of users’ emotional responses during voice user interface interactions. Their research results suggest that the physiological measure of facial expression and its extracted feature, automatic facial expression-based valence, is the most informative of emotional events lived through voice user interface interactions. By collecting and studying audio data from month-long deployments of the Amazon Echo in participants’ homes, Porcheron et al. [11] documented the methodical practices of VUI users and how that use is accomplished in the complex social life of the home. Based on the analysis of the collected data, they claimed that the response from the device is the primary ‘account’ of the system state and indicator of trouble and suggested a conceptual shift towards considering response

design as the design of interactional resources for users. Huang et al. [12] studied the differences between seven major “acoustic features” and their physical characteristics during voice interaction with the recognition and expression of “gender” and “emotional states of the pleasure-arousal-dominance (PAD) model.” They concluded that the gender and emotional states of the PAD model vary among seven major acoustic features. Moreover, their different values and rankings also vary. Bottaci et al. [13] developed software that links a selected text-to-speech (TTS) synthesizer with an automatic speech recognition (ASR) engine, producing a chatbot to explore the psychological implications of artificial speech emotion. By analyzing the data from their voice interaction experiments, they asserted that humans complain that a synthesizer sounds “robotic” or “alien” because the voice signal is expressing the wrong emotion, leading to confusion and miscommunication. Kim et al. [14] investigated the effect of nonverbal vocal cues in speech interaction on the user’s perception of the agent. They designed the experiment to analyze participants’ responses regarding intimacy, similarity, connectedness, enjoyment, and ease of use of the speech interaction agent. The study result showed that using nonverbal vocal cues on empathic feedback contributes to establishing an interpersonal relationship with the agent. Some studies [15], [18] have found that speech rate, which includes the speed at which words are spoken as well as the length of pauses and variations in speech flow, is one of the main acoustic contributors to the display of emotional speech. No research on user’s affection while waiting for the response from VUI has been found yet.

B. TIME PERCEPTION

In time psychology research, time perception is a sustained and sequential response made by individuals to time stimulation that directly affects their organs. That is to say; individuals can judge their perception of the duration and speed of things without the help of any timers. People’s perception of time is similar to that of colors, shapes, and temperatures, which is an instinct people are born with [19].

An Individual’s sense of time is altered by his emotions to such an extent that time seems to fly when we are having fun and drags when we are bored [20]. On the contrary, an individual’s perception of time affects his emotion correspondingly. For instance, when an individual perceives the duration of content loading as short, positive emotions might arise. To shorten pedestrians’ experienced waiting time, Cao et al. [21] explore how the tempo and pitch of audible pedestrian signals influence time estimation. Cao et al. [22] explored the motion design’s impact on users’ time perception when users are waiting to load APP pages. The result shows that the waiting time perception of APPs is related to the loading motion types, the combination type of loading motions can effectively shorten the waiting time perception. Chen et al. [23] investigated the influences of the loading’s present duration as well as loading’s type on time perception

and emotional experience. These findings indicated that time perception and emotional experience depended on the loading’s present duration and type; reducing the present duration and using the ‘video’ type of loading can influence the time perception and reduce the experience of anxiety. Increase the user experience. Noulhiane et al. [24] investigated the influence of emotions on timing in reproduction and verbal estimation tasks. The results show that emotion-induced activation increases pacemaker rate, leading to a longer perceived duration. Appelqvist-Dalton et al. [25] explored the effect of sensory modality, arousal, and valence on how participants estimate durations in a film which is a multimodal stimulus. Their research shows that clip durations were judged to be shorter than actual durations, and visual-only clips were perceived as longer (i.e., less distorted in time) than auditory-only and audiovisual clips.

Individual’s perception of time has been measured through three methods as follows:

- 1) The first method is the determined time estimation: after completing a designated task, each subject will give the determined duration as feedback by pressing specific keys or reporting it orally. For example, in a waiting time study on consumption scenes, Hui et al. asked the subjects to estimate their time perception orally after they had completed the task [26], [27].
- 2) As for the second method, the subjects are invited to compare the target stimulation time with the standard stimulation time. However, this method is rarely used compared to the determined time interval estimation. The abovementioned two methods are called by a joint name—time interval estimation, namely, the estimation of time intervals [28].
- 3) Regarding the third method, the subjects are asked to select the time lapsing speed using a scale. Compared with the first two methods, the scale method has higher consistency with indexes, such as the subjects’ time satisfaction. Therefore, it is better at reflecting the influence of the subjects’ time perception on their subsequent behaviors and attitudes [29].

To summarize, the influence on time perception is researched generally, from motion, the color of stimulus, and stimulus types to visual, auditory, and audiovisual modalities. However, the duration of waiting for the feedback from VUI is barely concerned.

III. VUI FEEDBACK TIME PERCEPTION EXPERIMENTS

A. SELECTION OF SUBJECTS

Research Report of China Enterprise Cases for Intelligent Voice Assistant appeared in 2019 and showed that Chinese users of intelligent voice assistants are relatively young. Over 53.0% were from 20 to 35 years old, and about 80.2% had bachelor’s degrees or above. Users first use intelligent voice assistants for information search, and function calls second [30]. Therefore, 40 teachers and students, 20 males and 20 females, aged between 25 and 30 years old, who

have normal hearing as well as standard and clear pronunciation of Mandarin without speech abnormality, from a senior high school were selected for the experiment of this paper. All experimental subjects have experience using VUI voice assistant applications on personal computers (PC) or mobile devices.

B. EXPERIMENTAL TASK FLOW

According to the different purposes for which users use voice interaction devices, the types of dialogue tasks can be divided into non-task-oriented dialogues and task-oriented dialogues [31], [32]. Non-task-oriented dialogues primarily refer to forms of interaction in which users have no clear expectations about the feedback from the voice interaction system. Typical application forms include listening to music, stories, and operas. In this scenario, the user’s purpose is to kill time and relieve loneliness, and they are not so eager to respond to the system. The feedback time is not so sensitive that it can influence the user experience [33].

Task-oriented dialogues mainly refer to a scenario that voice interaction systems assist users in accomplishing specific tasks, such as checking the weather, making hotel or restaurant reservations, et al., by single or multiple rounds of dialogues [34], [35]. When utilizing such functions, users activate the device and give the corresponding voice command, waiting for the responses from voice interaction systems. Then users extract the information needed from the feedback played by the device and store it in short-term or long-term memory, which can be applied to the specific task. In task-oriented dialogues, feedback time is a key factor that greatly affects the user experience [36]. Therefore, the Task-oriented dialogues were selected for the feedback time perception experiment.

Based on the results from the interviews with 40 users, “weather check,” “function call,” and “schedule planning” are the common interactions in daily use of voice interaction systems. Therefore, these tasks were selected to be executed in the experiment.

The experimental task flow is shown in Figure 2. Before the experiment, every experimental subject was required to get familiar with the experimental task flow using the voice assistant experimental platform. When the experiment was started, the experimental platform was loaded with a Google Chrome 64-digit browser and shown on an about 35-inch in-plane switching (IPS) screen. Experimental subjects were required to finish the voice dialogue task in a quiet lab, sitting directly in front of a screen, 30 cm apart. The specific progress of the dialogue task is as follows.

At first, subjects speak specific awakening words—“Hello, Xiaozhi” to VUI in a normal state of speech. Then, the experimental prototype judges the end of the speech. After the set feedback time, the prototype informs the subjects with voice and words “feedback” that VUI has been switched to the state of being awakened. When subjects realize that VUI is now awakened, they will give specific voice commands

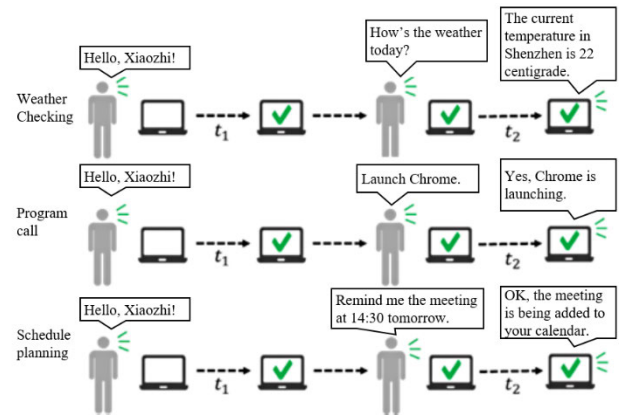


FIGURE 2. Human-computer Dialogue task flow.

to VUI. After the experimental prototype judges the end of their speech, VUI will respond to the subjects after the set feedback time with voice and word “feedback,” as shown on the screen. Experimental subjects need to finish three dialogue tasks, including “weather check,” “function call,” and “schedule planning.”

Task 1 requires Xiaozhi-the speech robot-to access the local weather information from the internet and translate it into speech and speak out. Task 2 requires Xiaozhi to call the application installed on the experimental PC, launch it, and then report the result of execution. Task 3 requires Xiaozhi to call the local application calendar, write the schedule planning into it, and then report the result of execution. These three tasks differ in execution steps and resources, which may affect the subjects’ waiting time expectations.

As shown in Table 1, 18 types of feedback time are set in the VUI experiment to study the effects of different feedback times on users’ time perception and the relationship between users’ time perception and speech rate. Every dialogue task’s feedback time is selected randomly, and each could not be used for the second time in the experiment. Therefore, each dialogue task should be completed 18 times, and each time experimental subjects had a 30-second break before the next one.

TABLE 1. VUI feedback time setting.

VUI feedback time setting (ms)					
150	350	550	750	950	1150
1350	1550	1850	2150	3150	4150
5150	6150	7150	8150	9150	10150

C. PLATFORM FOR EXPERIMENTAL

This study developed an experimental platform prototype for VUI-based Web voice assistant applications using HTML5, CSS3, and JavaScript, as shown in Figure 3.

A function or calculation expression was called after the designated millisecond (ms) through the “setTimeout ()” method in JavaScript, thus, realizing the user-defined

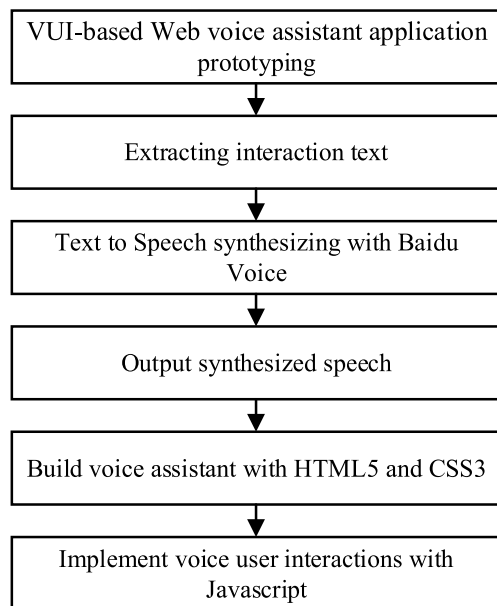


FIGURE 3. The prototype development process of the experimental platform.

function of VUI feedback time. The voice corpus to be developed and used was synthesized by the “Baidu speech synthesis system,” with the following parameters: standard female voicebank, a sampling rate of 48 kHz, and 32 bits (floating point).

To explore the influence of VUI feedback time on user’s experience and eliminate the disturbance from other factors, the following configuration is considered:

- 1) The variables were controlled in the VUI design for the experimental platform, and the filler content,
- 2) The operational difficulty level and function expectation of VUI in each dialogue task were fixed, ensuring that only the VUI feedback time was changed in each task.

The interactive prototype of the VUI experimental platform is shown in Figure 4.

IV. DATA COLLECTING METHODS

A. METHODS OF TIME PERCEPTION MEASUREMENT

In this study, the Feedback Time Perception (FTP) of the subjects was measured using the scale method. A five-point Likert FTP scale (Table 2) was designed. After completing each dialogue task, each subject was asked to select his/her time perception caused by this dialogue task’s feedback time from the five descriptions in this scale. Each subitem was converted into the corresponding score hereafter to quantify the subjects’ time perception according to the complementary relationship between each time perception description and the score, as seen in Table 2. The corresponding relationship between descriptive statements and scores in this scale is as follows: The higher the feedback time’s score, the better the time perception caused by this feedback time to the user will be.

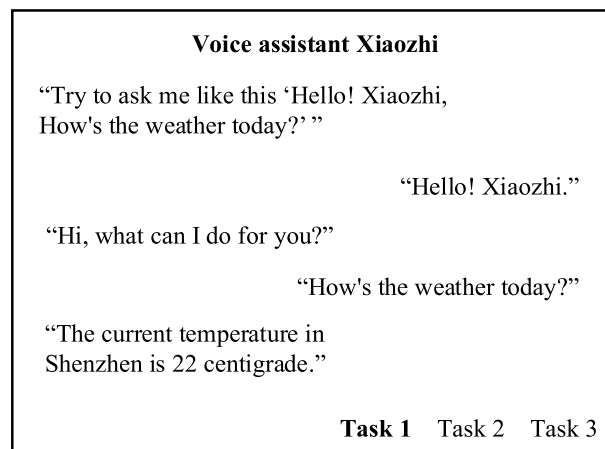


FIGURE 4. VUI interactive prototype of the experimental platform.

TABLE 2. Items and score distribution in FTP scale.

Dimension	Subitem	Description	Score
FTP	FTP1	The feedback time is too fast for me to accept	1
	FTP2	Being slightly fast, the feedback time is acceptable	3
	FTP3	The feedback time is just right	5
	FTP4	Being slightly slow, the feedback time is acceptable	3
	FTP5	The feedback time is too slow to accept	1

B. METHOD OF EMOTION QUANTIFICATION

Users’ satisfaction with a product is affected by their subjective feelings. Snyder and White’s study indicates that individuals will make positive choices and judgments for a thing with joyful emotions and negative ones with negative emotions [37]. A two-dimensional arousal-valence emotion space model [38], [40] was used to describe user emotions to investigate different VUI feedback time influences on users’ subjective emotions. After choosing the time perception given by the current VUI feedback time to themselves in the FTP scale, each subject was required to score their feelings from two aspects-valence and arousal-within the score 1-9.

C. METHOD OF SPEECH RATE EXTRACTION

The experimental tasks were executed using Adobe Audition CC2015, SGC-598 high-fidelity sound acquisition device, and Conexant Smart Audio HD sound card. While the experimental tasks were not affected, the subjects’ speech sounds were recorded in the experimental system background (audio format: 16 kHz, mono track, and 16 bits) to extract the subjects’ speech rates.

Related studies have shown that the Syllables Per Minute (SPM) and Words Per Minute (WPM) are two key indexes used to evaluate the speech rate. Neither of which is superior to the other, the two indexes present an extremely significant correlation, and anyone can be used to measure the

speech rate [41]. According to Chinese Phonology, a Chinese character is pronounced with a Chinese syllable, which generally includes an initial consonant and a vowel [42]. Two calculation methods are used in the SPM-based measurement of speech rate. The first method includes the pause time of silence (also called speed of sound) in the calculation, while the second one does not [43]. According to the study of Cao J. F., the auditory sense will be inconsistent with the subjective and objective evaluation of the speaker’s actual speech rate if the first method is used as the measurement criterion. After the pause time of silence is excluded, the statement will be in line with the actual auditory sense with the same number of syllables, along with a shorter total duration and higher speech rate [44]. As the number of syllables of voice command was required to be identical among all subjects in the experiment, SPM was used as the evaluation index for speech rate in this study. Moreover, it did not take a long time for each subject to give the voice command to the VUI, so the speech rate measurement method, which did not include the pause time of silence, was adopted.

By reference to the above measurement criteria for speech rate and combining the practical situation of the experimental tasks, the ratio of “syllables in the voice command-carrying statement (S)” to “total duration after the pause time of silence in the voice command-carrying statement spoken by the subject is excluded (T_w).” Namely, syllables per second (SPS) was used as the speech rate of the subject and denoted as V_s (unit: syllables/s). The pause time of silence in the subjects’ acquired corpora was automatically scanned and deleted using Adobe Audition software. Afterward, the syllables in a single statement (S) and the total duration of pause time deleted from the single statement (T) were extracted and processed through Equation (1) and Equation (2). Then the speech rate V_s of each subject was obtained.

$$T_w = T \times 10^{-3} \tag{1}$$

$$V_s = \frac{S}{T_w} \tag{2}$$

Following the automatic deletion of pause time indicated in Equation (1), the total duration of voice T (unit: ms) was converted into a total duration of T_w (unit: s). Afterward, T_w processed through Equation (1) was substituted into Equation (2) to obtain the speech rate V_s (unit : syllables/s) of each subject.

V. RESULTS AND ANALYSIS

A. INFLUENCE OF FEEDBACK TIME SETTING ON TIME PERCEPTION

During the experiment, every subject interacted with 54 materials, which are voice samples of 3 tasks with 18 feedback time configurations. After every human VUI interaction, subjects were asked to score (1-5) the material according to their perception of the feedback time of the material. The time perception scores of 40 subjects at different feedback times in the three tasks are summarized in Figure 5. As shown in the figure, the user’s time perception score was elevated with

the increase in the VUI feedback time (<750 ms), but the situation was the contrary when the VUI feedback time was longer than 750 ms. The single factor analysis of variance shows that no significant differences existed among the three different tasks in the time perception score. Therefore, the user perception of VUI feedback time was not affected by the different types of VUI voice interaction tasks.

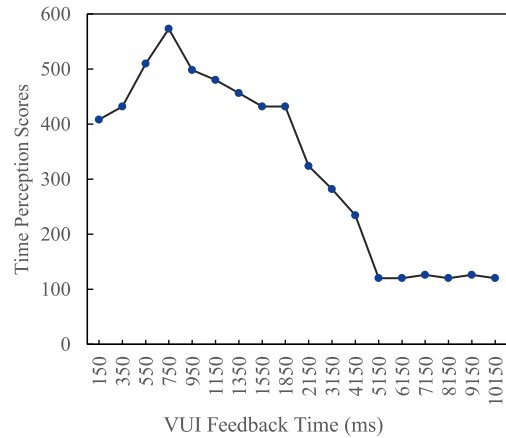


FIGURE 5. Feedback time-dependent change of subjects’ time perception.

The normality test of feedback time shows that it is not normally distributed. Then Spearman correlation between feedback time and perception time scores is verified. As is shown in Table 3, the Spearman correlation coefficient is -0.842 , which shows a strong negative correlation between the two variables. That is to say, when feedback time becomes longer, the user perception of the waiting time gets worse.

TABLE 3. Correlation between feedback time and Perception Time Scores.

		Time percep- tion scores	Feedback time
Spearma n's rho	Time per- ception scores	1.000	-.842**
	Correlation Sig. (2-tailed)	.	.000
	N	18	18
Feedback time	Correlation Sig. (2-tailed)	-.842**	1.000
	Correlation Sig. (2-tailed)	.000	.
	N	18	18

** . Correlation is significant at the 0.01 level (2-tailed).

The mean emotional arousal-valence space scores of 40 subjects within the feedback time of 150-10,150 ms were calculated and plotted in Figure 6. The x-coordinate displays the level of valence, namely, the unpleasant to the pleasant. In this case, one means unpleasant, and 9 represents pleasant. On the other hand, Y-coordinate refers to the arousal level, namely, the range from deactivation to activation. Again, 1, in this case, means deactivation, and 9 represents activation.

As shown in Figure 6, while the feedback time gets longer, subjects’ arousal and valence degree become lower. Feedback time shorter than 3150 can bring a positive effect, and longer than 3150 ms causes low valence levels, which means a negative effect. Subjects are active when feedback time is

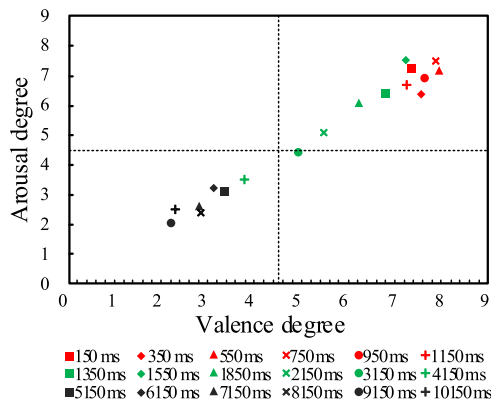


FIGURE 6. Two-dimensional spatial distribution of emotional valence at different time points.

shorter than 2150 and passive when feedback time is longer than 2150 ms.

Combining data in Figure 5, Figure 6, and the post-experimental user interview, four cases arise as follows:

- 1) When the VUI feedback time is 110-550 ms, the time perception scores ascend, and positive affect elicits. However, it was considered “too short” and caused a sense of hurry in the post-experimental user interview.
- 2) When VUI feedback time is within 550-1,850 ms, it was deemed “relatively appropriate,” which was enough to bring positive affect and satisfaction.
- 3) When VUI feedback time was within 1,850-4,150 ms, some subjects thought that “there is an obvious delay,” and others showed negative emotions during the waiting process.
- 4) When VUI feedback time was longer than 4,150 ms, all subjects thought that “it was too seriously delayed to accept” and showed low arousal and valence levels.

The experiment also demonstrates that users will be affected by different VUI feedback times during the VUI-based emotional experience of voice interaction. Therefore, according to the experimental results, different VUI feedback times influenced users’ time perception and subjective emotions.

The threshold time for users to wait for feedback during the VUI-based voice interaction process was 1,850 ms. The users had the best subjective feelings and psychological experience, with the highest score at the VUI feedback time point of 750 ms. In comparison with short feedback time (<750 ms), long feedback time (≥ 750 ms) exerted significantly different influences on users’ time perception and subjective emotions. To be more specific, when the VUI feedback time exceeded the threshold limit value of the waiting time, users would show emotions of low arousal and valence; and their acceptance level for feedback time was considerably lowered.

B. EXPLORATION OF CORRELATION BETWEEN TIME PERCEPTION AND SPEECH RATE

Descriptive Statistics, Pearson Correlation Analysis, and Linear Regression Analysis were carried out for the experimental data with IBM SPSS Statistics.

After the experiment, three groups of speech rate data V_s were acquired from each sample. An independent-samples t-test was performed for “ V_s ” and “gender” of the subjects. The results showed that the difference between “ V_s ” and “gender” was of no statistical significance ($\alpha = 0.559 > 0.05$), indicating an insignificant correlation between the two. This conclusion is consistent with Ha-Kyung K et al. results; namely, the speech rate is not significantly affected by gender or means of expression under independent circumstances [20]. The repeated test of variance showed that the speech rates of the subjects in completing the three different tasks were not significantly different; therefore, V_s was not influenced by different VUI voice interaction tasks. Therefore, 40 groups of V_s data were obtained by taking three groups of the mathematically expected value of V_s from each experimental sample, as shown in Figure 7.

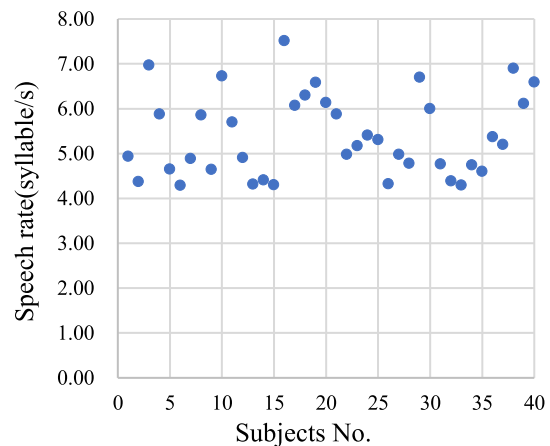


FIGURE 7. V_s distribution graph of subjects.

As verified in section IV-A, different VUI voice interaction tasks did not influence users’ time perception of VUI feedback time. The time point corresponding to the dimension FTP3 in the FTP scale contributed to the best time perception. Therefore, the mathematically expected values of all time points selected in the FTP3 column of the FTP scale by each subject were taken as the “optimal feedback time points.” The parameter V_s and each subject’s corresponding “optimal feedback time” were organized as shown in Figure 8.

The normality test of the optimal feedback time shows that it is not normally distributed. Then a non-parametric test is conducted on it. As is shown in Table 4, the correlation between speech rate and optimal VUI feedback time is verified, and the Spearman correlation coefficient is -0.742 , which shows a strong negative correlation between the

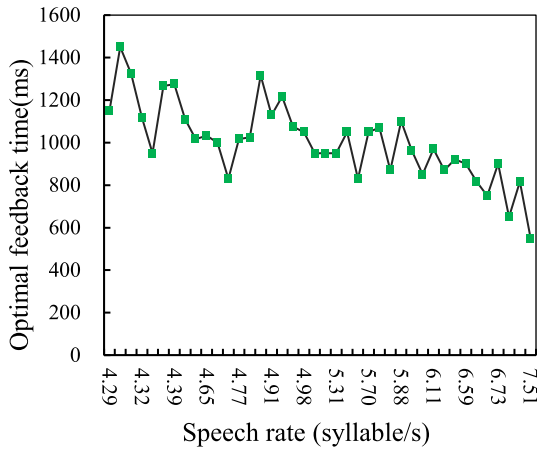


FIGURE 8. V_s -dependent change of subjects' optimal feedback time.

TABLE 4. Correlation between speech rate and optimal feedback time.

		Speech rate	Optimal feedback time
Spearman's rho	Correlation Coefficient	1.000	-.742**
	Sig. (2-tailed)	.	.000
	N	40	40
Optimal feedback time	Correlation Coefficient	-.742**	1.000
	Sig. (2-tailed)	.000	.
	N	40	40

** . Correlation is significant at the 0.01 level (2-tailed).

two variables. This indicates that users with higher speech rates expect shorter feedback times.

The significance of the regression equation was tested through the analysis of variance (ANOVA) [39], and the significance value of ANOVA was $p = 0.000 < 0.05$. This value indicates that the linear relation regression model established between the independent variable " V_s " and the dependent variable "optimal feedback time" is statistically significant. The linear regression equation's goodness of fit was further calculated for the linear regression model's fitting effect. The analysis showed that the correlation coefficient and the linear regression model coefficient of determination were $R = 0.905$ and $R^2 = 0.819$, respectively; the latter was adjusted as $R^2 = 0.814$, manifesting higher goodness of fit of the regression equation. In other words, the correlation between the independent variable "SPS" and the dependent variable "optimal feedback time" can be explained by this model to a great extent.

As the time points, depicted in Figure 8, fluctuated differently due to their different slope values k , the original regression equation was set as $f(V_s) = kV_s + 1,827.30$. Using this equation, the value range of slope k of the linear regression model for "optimal feedback time" can be obtained. More specifically, the V_s data and the corresponding optimal feedback time $f(V_s)$ of 40 subjects were substituted into the above equation to obtain the slope values k at all points given in Figure 8, followed by a normality test of the k values via

IBM SPSS Statistics [45]. The results showed that the k values followed a normal distribution ($p = 0.2 > 0.05$), and the confidence level of 95% was taken from the mean k value to obtain its value range as $-178.54 \leq k \leq -142.77$.

Equation (3) exhibits the linear regression model of "optimal feedback time" established by taking the independent variable "SPS" as V_s and the dependent variable "optimal feedback time" as $f(V_s)$.

$$\begin{cases} f(V_s) = kV_s + 1827.30 \\ -178.54 \leq k \leq -142.77 \\ V_s > 0, f(V_s) > 0, f(V_s) \subset N \end{cases} \quad (3)$$

According to this model, the speech rate will force, to a certain level, the user to choose the optimal VUI feedback time specifically manifested as follows: During the user voice interaction process with VUI under normal speech status, the user speaks at a faster inherent speech rate (high V_s value) is more inclined to accepting shorter VUI feedback time.

C. VERIFICATION OF CORRELATION BETWEEN TIME PERCEPTION AND SPEECH RATE

The usability of the linear regression model was experimentally verified in this study. First, the user groups of intelligent voice assistants in China were described according to the 2019 statistics entitled "Research Report on Enterprise Cases of China Intelligent Voice Assistant." A total of 20 teachers and students (10 males and 10 females, aged from 20 to 35) were selected from a university for this experiment. The subjects have the following characteristics: normal hearing, common and clear pronunciation of Mandarin, and no speech abnormality. Besides, all subjects can use PC and mobile VUI voice assistant-type applications and do not participate in the previous experiment.

The verification experiment was concretely implemented: The SPS data is collected when the present subject gives a voice command using the specified method in section 2.4.3. The V_s distribution of 20 subjects is processed through Equation (1) and Equation (2), as shown in Figure 9.

The V_s of each subject is processed via Equation (3), k value is taken as the fitted slope, namely, $k = k_i = -152.63$. Then, the theoretical value (unit: ms) of "optimal feedback time" corresponding to each subject is obtained. During the experiment, the theoretical value of each subject's "optimal feedback time" is set as the experimental platform's VUI feedback time. Before experimenting, the Google Chrome 64-bit browser is loaded into the experimental platform and displayed on a 27-inch IPS display. Each subject is asked to sit directly facing the display (spacing: about 30 cm) in a quiet laboratory environment and complete the designated voice dialogue task with the VUI experimental platform. After completing this task, the subject selects the time perception triggered by the current VUI feedback time in the FTP scale, the scale items, and the score distribution presented in Table 2.

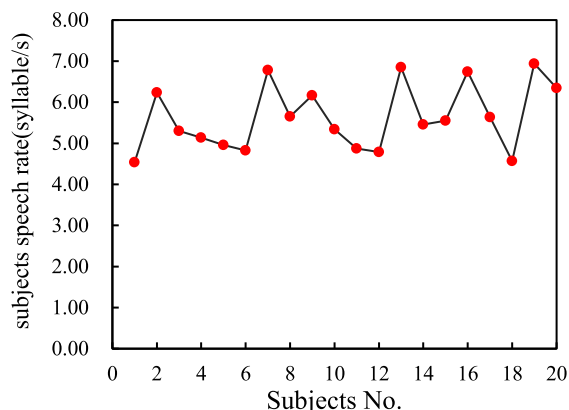


FIGURE 9. V_s distribution of subjects.

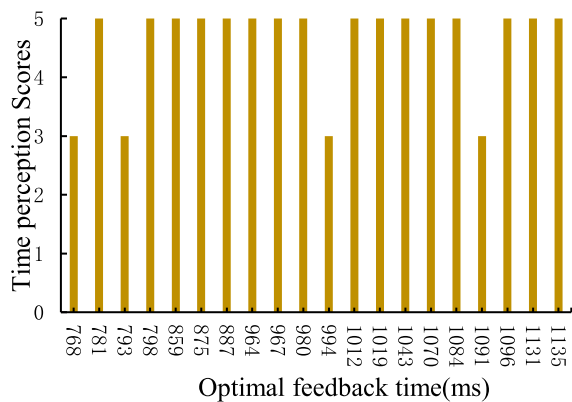


FIGURE 10. Time perception scores at optimal feedback time points.

The time perception scores of 20 subjects on their optimal VUI feedback time given by equation (3) are shown in Figure 10. It shows that 16 subjects were satisfied with the feedback time and 5 with the time perception score. Only four subjects gave 3, which also manifests a non-negative attitude toward the feedback time given by the proposed model.

Based on the IBM SPSS Pearson correlation analysis of 20-group “time perception” data and “optimal feedback time” data, no correlation is manifested between them ($p=0.531>0.05$). By combining Figure 10 and the post-experimental user interviews, it could be deduced that the feedback time acquired through the linear regression equation is the theoretical value of “optimal feedback time” acceptable by each user, with a fixed influence on user time perception. Although such theoretical value varied from user to user, this did not influence the users to acquire the time perception. It means “the feedback time is suitable” from the theoretical value of “optimal feedback time.” Therefore, users can acquire a good feedback experience in the VUI voice interaction process by setting the linear regression model “optimal feedback time” as the VUI feedback time.

VI. CONCLUSION

This paper discussed the time perception and user experience during VUI user interaction at different VUI feedback times. A VUI-based Web voice-assistant application was used as

the experimental prototype to investigate the influence of feedback time on user time perception during the VUI voice interaction using different VUI feedback times. The two-dimensional arousal-valence emotion space distribution was used to describe the influences of varying VUI feedback time on user emotional changes. Also, users’ speech rate data was collected during the experiment. The following conclusions were drawn:

- 1) Users’ time perception and subjective emotions are differently influenced by different VUI feedback times. 750 ms is the optimal VUI feedback time point at which the best users’ subjective feelings and psychological experiences are reached. The threshold limit time spent by users in waiting for the VUI feedback is 1,850 ms. If the VUI feedback time exceeds this value, it leads to user emotions with low levels of arousal and valence.
- 2) SPS of each user presents a significant negative correlation with the good optimal VUI feedback time point. When the users have a faster inherent speech rate under normal speaking status, it is easier for them to accept shorter VUI feedback time during the voice interaction process. The author has established a simple linear regression model based on the linear regression analysis results. The model verification experiment indicated that the proposed model is feasible.
- 3) It is also suggested that a user speech rate detection module can be added to the existing VUI voice interaction products or during the research and development of voice interaction products. Besides, the VUI feedback time can be adjusted according to the proposed model for a better user experience.

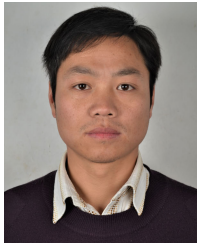
Influencing people’s perception of waiting time through visual cues can indeed effectively improve user experience in the process of human-computer interaction [21], [24]. However, they are expedient and remedial measures. The fundamental means to solve the waiting time problem is to set a waiting time according to each person’s different perception of time. We study the corresponding feedback time configuration in accordance with the user’s speech rate and propose a linear regression model of “optimal feedback time” which are established on the experimental statistics.

We believe that the research conclusions will effectually guide the voice interaction design research. Notwithstanding uncertain factors like user’s habits, emotions, cultural background, and cognitive ability will influence users’ satisfaction with the voice interaction experience in various pattern, the feedback time suggestions based on the study results still provide a reference for the R & D of the existing voice interaction products.

REFERENCES

[1] C. Pearl, *Designing Voice User Interfaces: Principles of Conversational Experiences*. Sebastopol, CA, USA: O’Reilly Media, 2016, pp. 103–115.
 [2] Z.-P. Lu and Y.-K. Dong, “The study based on usability test of voice interface in vehicle navigation,” *J. Packag. Eng.*, vol. 8, pp. 28–34, Aug. 2013, doi: 10.19554/j.cnki.1001-3563.2013.08.008.

- [3] R. Levitan, Š. Benuš, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *Proc. AAAI Spring Symp. Ser.*, 2015, pp. 1–8.
- [4] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *J. Phonetics*, vol. 38, no. 4, pp. 555–568, Oct. 2010.
- [5] A. D. MacIntyre and K. S. Scott, "Listeners are sensitive to the speech breathing time series: Evidence from a gap detection task, cognition," *Cognition*, vol. 225, pp. 1–30, Aug. 2022, doi: [10.1016/j.cognition.2022.105171](https://doi.org/10.1016/j.cognition.2022.105171).
- [6] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smart-phone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018, doi: [10.1109/ACCESS.2018.2800728](https://doi.org/10.1109/ACCESS.2018.2800728).
- [7] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1181–1185, Aug. 2018, doi: [10.1109/LSP.2018.2811740](https://doi.org/10.1109/LSP.2018.2811740).
- [8] Z. Ali and M. Talha, "Innovative method for unsupervised voice activity detection and classification of audio segments," *IEEE Access*, vol. 6, pp. 15494–15504, 2018, doi: [10.1109/ACCESS.2018.2805845](https://doi.org/10.1109/ACCESS.2018.2805845).
- [9] X. Tan and X.-L. Zhang, "Speech enhancement aided end-to-end multi-task learning for voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6823–6827, doi: [10.1109/ICASSP39728.2021.9414445](https://doi.org/10.1109/ICASSP39728.2021.9414445).
- [10] D. Swoboda, J. Boasen, P.-M. Léger, R. Pourchon, and S. Sénécal, "Comparing the effectiveness of speech and physiological features in explaining emotional responses during voice user interface interactions," *Appl. Sci.*, vol. 12, no. 3, p. 1269, Jan. 2022, doi: [10.3390/app12031269](https://doi.org/10.3390/app12031269).
- [11] M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples, "Voice interfaces in everyday life," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–12, doi: [10.1145/3173574.3174214](https://doi.org/10.1145/3173574.3174214).
- [12] K.-L. Huang, S.-F. Duan, and X. Lyu, "Affective voice interaction and artificial intelligence: A research study on the acoustic features of gender and the emotional states of the PAD model," *Frontiers Psychol.*, vol. 12, May 2021, Art. no. 664925, doi: [10.3389/fpsyg.2021.664925](https://doi.org/10.3389/fpsyg.2021.664925).
- [13] S. Pauletto, B. Balentine, C. Pidcock, K. Jones, L. Bottaci, M. Aretoulaki, J. Wells, D. P. Mundy, and J. Balentine, "Exploring expressivity and emotion with artificial voice and speech technologies," *Logopedics Phoniatrics Vocol.*, vol. 38, no. 3, pp. 115–125, Oct. 2013, doi: [10.3109/14015439.2013.810303](https://doi.org/10.3109/14015439.2013.810303).
- [14] J. Kim, W. Kim, J. Nam, and H. Song, "'I can feel your empathic voice': Effects of nonverbal vocal cues in voice user interface," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–8, doi: [10.1145/3334480.3383075](https://doi.org/10.1145/3334480.3383075).
- [15] K. Scherer, "Personality markers in speech," in *Social Markers in Speech*, K. Scherer and H. Giles, Eds. London, U.K.: Cambridge Univ. Press, 1979, pp. 147–210.
- [16] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1097–1108, 1993.
- [17] T. Bowles and S. Pauletto, "Emotions in the voice: Humanising a robotic voice," in *Proc. 7th Sound Music Comput. Conf.*, Barcelona, Spain, Jul. 2010, pp. 21–24. [Online]. Available: <http://smcnetwork.org/files/proceedings/2010/30.pdf>
- [18] S. Pauletto and T. Bowles, "Designing the emotional content of a robotic speech signal," in *Proc. 5th Audio Mostly Conf., Conf. Interact. With Sound*, Sep. 2010, pp. 1–8. [Online]. Available: <http://dl.acm.org>
- [19] A. Cooper, "The inmates are running the asylum," in *Proc. Software-Ergonomie Design von Informationswelten, Gemeinsame Fachtagung des German Chapter ACM*, Walldorf, Germany: ACM, 1999, pp. 8–11.
- [20] S. Droit-Volet and W. H. Meck, "How emotions colour our perception of time," *Trends Cognit. Sci.*, vol. 11, no. 12, pp. 504–513, Dec. 2007.
- [21] Y. Cao, X. Zhuang, and G. Ma, "Shorten pedestrians' perceived waiting time: The effect of tempo and pitch in audible pedestrian signals at red phase," *Accident Anal. Prevention*, vol. 123, pp. 336–340, Feb. 2019.
- [22] H. Cao and X. Hu, "Research on motion design for APP's loading pages based on time perception," *AIP Conf. Proc.*, vol. 1955, no. 1, 2018, Art. no. 040075.
- [23] D. Chen, K. Yao, X. tan, and Z. Yang, "The user experience of loading design: The influences of the loading's present duration as well as type on waiting time perception and emotional experience," *Chin. J. Ergonom.*, vol. 21, no. 4, pp. 6–12, 2015.
- [24] M. Noulhiane, N. Mella, S. Samson, R. Ragot, and V. Pouthas, "How emotional auditory stimuli modulate time perception," *Emotion*, vol. 7, no. 4, pp. 697–704, 2007, doi: [10.1037/1528-3542.7.4.697](https://doi.org/10.1037/1528-3542.7.4.697).
- [25] M. Appelqvist-Dalton, J. P. Wilmott, M. He, and A. M. Simmons, "Time perception in film is modulated by sensory modality and arousal," *Attention, Perception, Psychophysics*, vol. 84, no. 3, pp. 926–942, Apr. 2022, doi: [10.3758/s13414-022-02464-9](https://doi.org/10.3758/s13414-022-02464-9).
- [26] M. K. Hui and D. K. Tse, "What to tell consumers in waits of different lengths: An integrative model of service evaluation," *J. Marketing*, vol. 60, no. 2, pp. 81–90, 1996, doi: [10.1177/002224299606000206](https://doi.org/10.1177/002224299606000206).
- [27] M. K. Hui, M. V. Thakor, and R. Gill, "The effect of delay type and service stage on consumers' reactions to waiting," *J. Consum. Re-Search*, vol. 24, no. 4, pp. 469–479, Mar. 1998, doi: [10.1086/209522](https://doi.org/10.1086/209522).
- [28] B. Cuihua, C. Youguo, and H. Xiting, "The impact of number and numerical presentation mode on duration estimation," *Stud. Psychol. Behav.*, vol. 8, no. 3, pp. 161–165, Mar. 2010.
- [29] S. W. Brown, "Time, change, and motion: The effects of stimulus movement on temporal perception," *Perception Psychophys.*, vol. 57, no. 1, pp. 105–116, Jan. 1995, doi: [10.3758/bf03211853](https://doi.org/10.3758/bf03211853).
- [30] *Research Report of China Enterprise Cases for Intelligent Voice Assistant Appeared in Year 2019*, Shanghai, China: Iresearch Inc., 2018. [Online]. Available: <http://www.iresearchchina.com/>
- [31] H. L. Garcia, R. M. Gonzalez, H. G. Rosales, J. C. Padilla, C. G. Tejada, F. E. L. Monteagudo, C. A. Collazos, and A. M. Gonzalez, "Mental models associated to voice user interfaces for infotainment systems," *DYNA*, vol. 93, no. 1, p. 245, May 2018.
- [32] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 2, pp. 25–35, Dec. 2017.
- [33] M. Snyder and P. White, "Moods and memories: Elation, depression, and the remembering of the events of one's life," *J. Personality*, vol. 50, no. 2, pp. 149–167, Feb. 1982, doi: [10.1111/J.1467-6494.1982.TB01020.X](https://doi.org/10.1111/J.1467-6494.1982.TB01020.X).
- [34] R. Ziman and G. Walsh, "Factors affecting seniors' perceptions of voice-enabled user interfaces," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–6.
- [35] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 2, pp. 25–35, Dec. 2017.
- [36] Q. Ma, R. Zhou, C. Zhang, and Z. Chen, "Rationally or emotionally: How should voice user interfaces reply to users of different genders considering user experience?" *Cognition, Technol. Work*, vol. 24, no. 2, pp. 233–246, May 2022.
- [37] B. Stigall, J. Waycott, S. Baker, and K. Caine, "Older adults' perception and use of voice user interfaces: A preliminary review of the computing literature," in *Proc. 31st Austral. Conf. Hum.-Comput.-Interact.*, Dec. 2019, pp. 423–427.
- [38] R. Tato, R. Santos, and R. Kompe, "Emotional space improves emotion recognition," in *Proc. 7th Int. Conf. Spoken Lang. Process.*, 2002, pp. 2029–2032.
- [39] R. Cowie, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001, doi: [10.1109/79.911197](https://doi.org/10.1109/79.911197).
- [40] B. Xie, "Research on key issues of Mandarin speech emotion recognition," Ph.D. dissertation, College Comput. Sci., Zhejiang Univ., Hangzhou, China, 2006.
- [41] K. Ha-Kyung, "The speech rate in monologue and reading in normal adults," *J. Audiol. Speech Pathol.*, vol. 23, no. 3, pp. 240–243, Mar. 2015.
- [42] Li Wang, *Chinese Phonology*. Beijing, China: Zhonghua Book Company, 2014, pp. 93–110.
- [43] L. Yinghao and K. Jiangping, "Effects of speech rate on segment production in Standard Chinese," *J. Tsinghua Univ. Sci. Technol.*, vol. 57, no. 9, pp. 963–969, Sep., 2017, doi: [10.16511/j.cnki.qhdxxb.2017.26.048](https://doi.org/10.16511/j.cnki.qhdxxb.2017.26.048).
- [44] J. Cao, "Characteristics and changes of speech speed," in *Proc. 6th Chin. Academic Conf. Modern Phonetics*, 2003.
- [45] Z. Minjing, L. Ya'na, and X. Zhiqun, "Linear regression equation study of relevant inspection question," *Value Eng.*, vol. 31, no. 2, pp. 1–2, Feb. 2012, doi: [10.14018/j.cnki.cn13-1085/n.2012.02.001](https://doi.org/10.14018/j.cnki.cn13-1085/n.2012.02.001).



JUNFENG WANG was born in Dali, Shaanxi, China, in 1981. He received the B.S. degree in industrial design from the Southwest University of Science and Technology, in 2004, and the M.S. and Ph.D. degrees in industrial design from Northwestern Polytechnical University, in 2007 and 2016, respectively. From 2007 to 2019, he was a Lecturer and an Associate Professor with the Southwest University of Science and Technology. Since 2019, he has been an Associate Professor with the College of Design and Innovation, Shenzhen Technology University. He is the author of 13 books, more than 20 articles, and more than 20 inventions. His research interests include human system interaction in cyber physical social systems, experience and ergonomics in voice user interface, and service provision of the intelligent Internet of Things.



SHIYU DONG received the B.E. degree from the Department of Industry Design, Southwest University of Science and Technology, in 2020. She is currently pursuing the M.E. degree with the College of Design and Innovation, Shenzhen Technology University. Her research interest includes human-machine interaction in automobile field.



YUE LI received the B.S. degree in information and control engineering and the M.S. degree in industrial design engineering from the Southwest University of Science and Technology, in 2017 and 2020, respectively. In 2020, he was an Interaction Designer with the China General Nuclear Power Group. His research interests include human-computer interface experience, voice user interface design and optimization, and data visualization.



SHUYU YANG received the B.E. degree from the Department of Industrial Design, Sichuan University of Science and Engineering, in 2017. She is currently pursuing the M.E. degree in industrial design engineering with the School of Manufacturing Science and Engineering, Southwest University of Science and Technology. Her research interest includes human-computer interaction in intelligent voice systems.



JIALIN LI received the B.E. degree from the College of Astronautics, Nanjing University of Aeronautics and Astronautics, in 2019. She is currently pursuing the M.E. degree with the College of Design and Innovation, Shenzhen Technology University. Her research interests include human-computer interaction and user experience.

...