



The Unit and Size of Information Supporting Auditory Feedback for Voice User Interface

Min Chul Cha, Hyo Chang Kim & Yong Gu Ji

To cite this article: Min Chul Cha, Hyo Chang Kim & Yong Gu Ji (2023): The Unit and Size of Information Supporting Auditory Feedback for Voice User Interface, International Journal of Human-Computer Interaction, DOI: [10.1080/10447318.2023.2179214](https://doi.org/10.1080/10447318.2023.2179214)

To link to this article: <https://doi.org/10.1080/10447318.2023.2179214>



Published online: 19 Feb 2023.



Submit your article to this journal [↗](#)





View related articles [↗](#)



View Crossmark data [↗](#)

The Unit and Size of Information Supporting Auditory Feedback for Voice User Interface

Min Chul Cha^a , Hyo Chang Kim^b , and Yong Gu Ji^a 

^aDepartment of Industrial Engineering, Yonsei University, Seoul, South Korea; ^bStanford Center at the Incheon Global Campus, Stanford University, Seoul, South Korea

ABSTRACT

The purpose of this study was to explore the unit of information and the size of the unit for designing a voice user interface. Through two experiments, this study investigated what form the information (the unit of information) should take and what size of that (the size of unit) should be when people were provided information by voice interfaces. Participants were presented with a task to recall (OX quiz) by listening to and remembering information (based on an encyclopedia) provided by smart speakers. In Experiment 1, it was revealed that participants stored information in their memory span on a sentence-by-sentences basis to determine how much information they could remember. In Experiment 2, sentence-based information was presented in various sizes, and participants evaluated 17 information units consisting of up to nine words as their memory limit. This information unit-based voice interface design could help improve users' memory performance and usability.

1. Introduction

Conversations with Voice User Interface (VUI) systems using AI are no longer special. In 2020, there were 87.7 million people in the United States with at least one smart speaker (Voicebot.ai, 2020b) and 129.7 million people with in-car voice assistants (Voicebot.ai, 2020a). Smart speakers can facilitate users' ability to get information (news, weather, encyclopedia, etc.), play music, and control household devices regardless of place (home, car, office, etc.).

These VUI systems are used to replace previous interfaces due to their input speed, hands-free, and intuitive advantages (Pearl, 2016). The advantages of VUI are expected to reduce the cognitive load on users and VUI is being embedded in various devices, such as cars, computers, smartphones, TVs, etc. VUI is known to reduce repetitive tasks and mental workload in web search and driving (Kim et al., 2019; Lee & Ji, 2019; Watanabe et al., 2007), and thus has a low level of effort or a low risk of accidents. In addition, in the study of Bickmore et al. (2009), information delivery through voice interfaces helped patients with low health literacy and showed high levels of satisfaction in the medical healthcare context.

However, according to the multiple resource model by Wickens (2002), providing additional information to users with other modalities that are not in use, rather than the same modality, allows users to process the information in parallel and reduces the risk of multitasking. Therefore, providing information to the driver via auditory rather than

visual can reduce these risks and the advantages of VUI can be utilized. Studies to utilize VUI or audio-voice systems already showed that users had improved performance and user experience in VUI compared to previous systems (Perlman et al., 2019; Strayer et al., 2019). They revealed that the voice interface minimizes the effects of driver distraction and generates better driving performance (lane keeping, quicker response time) and fewer safety-critical events compared to using manual controls and visual displays (Peissner & Doebler, 2011). Even in the performance of secondary tasks and user experience, the voice interface showed better results (Lo & Green, 2013; Yager, 2013).

Conversely, some studies reported the disadvantages of VUI. For instance, in the past, speech recognition technology had a high error rate and made it difficult for users to input commands at once. This error rate led to people spending more time and effort to correct errors (Hauptmann & Rudnick, 1990). Also, due to the specific form of the command, it required learning and caused low usability (Hua & Ng, 2010). However, these issues are being addressed by advances in speech recognition and natural language processing technologies, and some studies have been conducted to improve the usability by developing VUI command methods (such as "Barge-in" and "out-of-turn") and error recovery methods (Kim et al., 2019; Mane et al., 1996; Perugini et al., 2007; Yankelovich et al., 1995).

Despite technological advances in VUI, the nature of speech requires users to rely on linear and ephemeral short-term memory to recognize auditory output (Jung et al.,

2020). Depending on the user's situation, VUI is often used in systems without displays or in multitasking. Under those circumstances, users can not perceive visual information and rely solely on auditory information provided by the systems. As mentioned earlier, the linear nature of auditory information causes users to lose in the menu structure and places heavy demands on short-term memory (Howell et al., 2006). Aylett et al. (2014) and Shneiderman (2000) also have reported that a cognitive load is required for remembering items on audio-only output systems. If a large amount of auditory information is continuously presented, users have to listen to the system for a long time, which may cause a high workload (Winsum et al., 1999). Therefore, to reduce the negative effect of voice information on the user, research is needed to refine the information so that the user can easily process it.

This study aims to explore how to refine information and deliver it to users to minimize the cognitive load caused by the auditory information provided by VUI. Using encyclopedia information, the unit and size of information perceived by users when users interact with VUI were tested and analyzed. In the first experiment, which of the three information units (letters, words, and sentences) participants used for information processing was investigated. The goal of the second experiment was to examine the size of the information unit and the potential benefit of extension of the information unit on information processing. The result of this study could support the natural and comfortable conversation with the VUI system by providing the user with the proper size of information in the proper form.

2. Background

The more difficult the information is to understand, the more cognitive requirements would demand from the user. Therefore, the proper presentation of the information contributes to users' memory. There are two views of the proper presentation of the information: the form and amount of information. We focus on these two perspectives to design the auditory feedback provided by voice interfaces, and a detailed literature review of each perspective is described below.

2.1. Voice user interface

Voice user interface is a conversational interface in which a user interacts with a system via spoken language (Cohen et al., 2004). VUI system recognizes the user's words, interprets them, finds answers to the user's requirements, and presents results through voice again. In the Human-Computer Interaction domain, VUI is related to intuitive and natural human behaviors (López et al., 2018). This is considered the most natural interaction method for operating a system, and its importance has been highlighted due to recent advances in artificial intelligence and robot technology (Amershi et al., 2019). Voice interfaces are no longer just manipulating state-of-the-art technologies such as robots or smartphones but are gradually becoming a way of

manipulating past technologies such as refrigerators, washing machines, and lights (Sawan et al., 2013).

One of the primary characteristics of VUI is that the modality is auditory. Cohen et al. (2004) said that VUI that use auditory interfaces further challenge human memory and attention because they present information serially and non-persistently. Thus, they also insisted that the cognitive load of users should be minimized, and presented several cognitive challenges to solve them. One of them was the minimization of memory load by adjusting the menu size of information. In studies on human memory, it was revealed that people naturally divide items into several groups and that recall is best when receiving information in similar forms (Dirlam, 1972; Glanzer & Razel, 1974; Wickelgren, 1964). Therefore, in this study, it was expected that the cognitive load caused by the voice interface could be reduced by adjusting the unit and size of information, and this was explored through two experiments.

2.2. Unit of auditory information

Information provided from voice interfaces is immediately processed and stored in the short-term memory span using working memory (Baddeley, 2000). By investigating which unit of information is processed or stored at this time, this study attempted to find out the form of information suitable for users. Comparing performance on lists that differ in the length of their units is the general strategy for determining the unit size (Glanzer & Razel, 1974). Although a series of words is generally presented to a subject in conventional research, the sequence processed as a unit in short-term storage might be at any one of the following levels: (1) letters; (2) words (these would include multiple morpheme sequences such as farmhouse); (3) sentences and (4) units larger than sentences. Although morphemes exist as a unit with minimal meaning in other languages, there was a severe inter-morpheme coarticulation problem that could be raised due to short morphemes in Korean (Kwon & Park, 2003). Therefore, since the analysis unit should be defined and used as a word in which several morphemes are merged in Korean, and the morpheme unit was excluded from this study.

Early studies on memory employed the sequence of numbers or letters, but recently, studies using words or sentences as units are also being conducted. Humans tend to segment long and complex sequences into shorter and simple sequences, called "chunks," which are easy to process and store. According to Gobet et al. (2001), the chunking mechanism creates links between meaningful nodes, and these links improve users' memory span. Although the chunking occurs automatically and helps increase the users' memory, it remains unclear how it enables the chunking of complex sequences (Asabuki et al., 2018).

In Experiment 1, the exam was conducted to identify the unit of information processed by users via chunking and define the appropriate form of auditory feedback accordingly. When users use chunking to process the auditory feedback, It was expected that letters, words, or sentences

would serve as a unit of information. Usually, letters or words are segmented or incomplete information, but sentences are complete information in themselves. Therefore, rather than letters or words, sentences were expected to be used for chunking.

2.3. Size of auditory information

Information size is a very important consideration to convey information to users. Information (e.g., precautions) that is required by users or must be provided to users sometimes contain too much information. In that case, numerous pieces of information exceed the users' memory capacity, and it is not only unremembered by users, but also becomes a burden to users or causes to give them a negative impression of the system (Ranney et al., 2005; Saariluoma and Jokinen, 2014; Watanabe et al., 2007).

Determining what size of information to provide begins with determining the size of the information unit. Therefore, to adjust the amount of information, the size of the information unit needs to be limited and the amount of information must be determined based on this. Previous researchers believed that the magical number 7 ± 2 was the limit of working memory (Miller, 1956). On the other hand, Baddeley and Hitch (1974) proposed 2s limits for a phonological rehearsal loop, and they implied that chunks could extend beyond the words. However, as opposed to this, Cowan (2001) proposed the small size of chunks, the magical number 4. Although there have been discussions on the various size of chunks, they had the common result that the size of the memorizing item could be expanded through chunking. Therefore, to understand the form and amount of information that users need, it was necessary to study not only the unit in which chunking occurs but also the size of the unit affected by the chunking mechanism.

The object of Experiment 2 was to figure out the size of the information unit that users can process at once and the maximum size of auditory information. Similarly, in the study of Gilchrist et al. (2009), which investigated the user's memory capacity, they used the number of sentences and the number of words that consist of those sentences to evaluate the participant's recall performance. Based on this previous study, an information unit was constructed with sentences of up to nine words, and an information block with multiple information units was presented to participants. The experiment was conducted to analyze the effect of the size of the information block on the amount of information stored by participants.

3. Experiment 1

Experiment 1 tested to identify the unit of information that was matched to the user's information processing. Because sentences are complete information, unlike letters or words, we predicted that the sentence would serve as a unit of information.

To investigate our hypothesis, we used the encyclopedia information on three topics (Dwarf pine, Dingo, and

Platypus). Participants were asked to listen to the auditory information on one topic from a smart speaker and to remember as much information as possible. The number of letters, characters, and sentences that they listened to was calculated and analyzed to figure out which one can be used as units of information.

3.1. Methods

The Experiment consisted of two parts: the pilot study, which was to confirm that the levels of difficulty of three topics are the same, and the main study. In the pilot study, participants listened to all information about one of three topics and took an OX quiz about what they heard. On the other hand, in the main study, participants listened to as much information as they wanted to remember. Figure 1 shows the experiment room setting, which includes the location and connection status of each device. To convince participants that a smart speaker was working on its own, the experimenter in a separate room controlled the speaker and sent information. The research was deliberated and approved by the Institutional Review Board (IRB No. 7001988-202009-HR-787-05).

3.1.1. Participants

Forty-eight participants for the pilot study (24 males and 24 females) and sixty-five participants for the main study (33 males and 32 females) from the local university were recruited. They ranged in age from 20 to 35 years old (Pilot: mean = 24.4, $SD = 2.91$ Main: mean = 24.9, $SD = 2.78$). All participants had experience using VUI systems. They participated in the experiment for about 20 min and were paid \$10.

3.1.2. Apparatus and stimuli

To deliver the information, a smart speaker, Clova (Naver Corp., Seoul, Korea), was used. It is widely used in Korea. The smart speaker was connected to the experimenter's computer using Bluetooth. There was a hidden microphone in the test room for listening to participants' voices.

Three topics (Dwarf pine, Dingo, and Platypus) in the encyclopedia were selected as stimuli, i.e., auditory information to be presented to participants. Those were the information that is not for experts but is unfamiliar to the public. Each topic consisted of 20 sentences, but the total number of letters and words was different (Dwarf pine: 298 letters, 89 words; Dingo: 413 letters, 114 words; Platypus: 441 letters, 138 words). There is an example of "Dingo" in Table 1, which was translated into English.

3.1.3. Pilot study

A pilot study was needed to see whether the difficulties of the three topics are the same or not. Participants were asked to listen to whole information (20 sentences) about one of three topics in Korean. The same information as the main

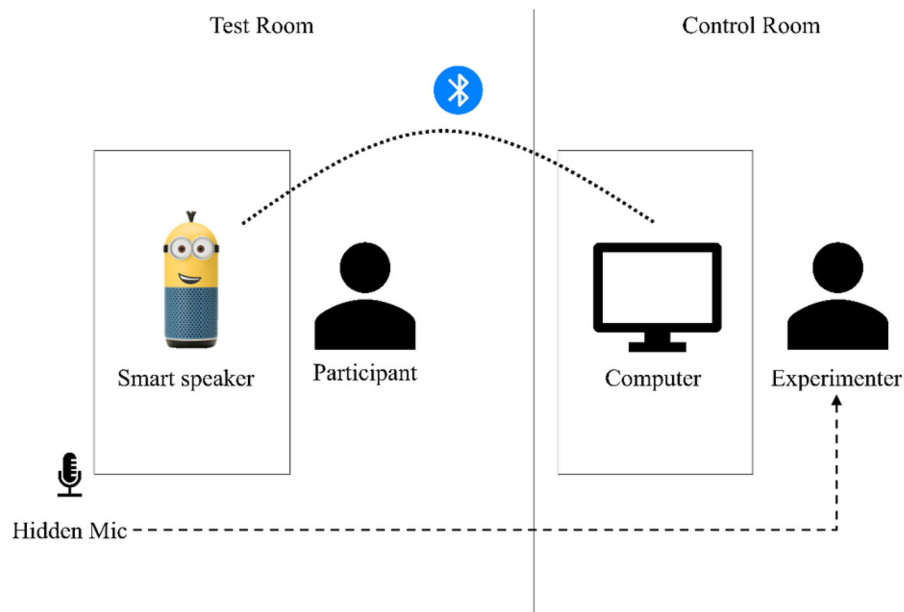


Figure 1. Experimental room setting. The experimenter controlled the smart speaker in a separate room by Bluetooth.

Table 1. Example of auditory information of “Dingo” in Experiment 1.

No	Sentence (translated from Korean to English)	Letters (in Korean)	Words (in Korean)
1	Dingoes are also called the Australian Wild Dog.	15	3
2	The length is 86–100 cm.	16	3
3	Their tail is 26–36 cm.	16	3
4	Their shoulder height is 44–64 cm.	16	3
5	The length of the ears is 9.5–10.5 cm.	18	3
6	They weigh 12–24 kg, slightly smaller than wolves.	22	6
7	Four legs are relatively long and the tail is tufted	18	7
8	They have a wide nose.	6	2
9	The ears are big and straight.	9	3
10	The colors are various (ginger, tan, white, and black), but tan is the main color.	33	7
11	It is believed that dogs that came to Australia from India or Southeast Asia 3500–4000 years ago turned into wild animals.	49	13
12	Alone or in pairs, sometimes in small groups.	19	8
13	Over the years, adapting to Australia’s harsh environment has made them very aggressive and rough.	41	11
14	They usually prey on rabbits, sometimes attacking sheep.	20	7
15	The breeding season is in the winter.	12	3
16	The gestation period is 10–12 weeks.	15	4
17	The size of the litter can range from one to 8 pups.	17	7
18	Females become sexually mature at the age of 2 and males at the age of 1–3.	25	7
19	Aboriginal people train pups and use them to catch lizards and snakes.	31	8
20	Dingoes are sometimes kept as pets.	15	6

study was presented to the participants in the pilot study. After listening, they took the OX quiz about the topic.

In the pilot study, the correct rate of OX quizzes was analyzed, and there was no significant difference between the difficulty levels of the three topics ($p = .166$). The average correct rate was 62.8% (see Table 2), and it was significantly different from the expected correct rate of OX quiz (50%) ($p < .000$). Thus, all topics had the same difficulties.

3.1.4. Procedure

The procedure of the study was basically as same as the pilot. The generic diagram of all experiments and the specific procedure of Experiment 1 was given in Figure 2. Participants were welcomed and then informed of the

contents of the study. The experimenter randomly assigned one of three topics to participants, and they were requested to ask the topic from the smart speaker in front of them in Korean. The participants were given another topic for training and they tried interacting with the smart speaker. When participants spoke the voice command, the experimenter sent the information in Korean to the smart speaker via a text-to-speech program. Then, the TTS program automatically spoke the information line by line. While listening to information, participants were also asked to say “stop” to stop listening when they realized their memory limitations. After listening, they took the OX quiz on the topic they listened to. As an example of the OX quiz, the question “Dingoes are bigger than wolves” was presented for the information “They are 12–24 kg, slightly smaller than

wolves.” Then the participants answered the question with “O” and “X.”

3.2. Results

3.2.1. Correct rate

Participants were randomly assigned to three topics: “Dwarf Pine,” “Dingo,” and “Platypus” with 24, 18, and 23 participants, respectively. Like the pilot study, the correct rate of the OX quiz was analyzed by topics. Table 2 showed the results of both studies. The ANOVA was performed to compare the effect of topics on the correct rate. There was no significant difference between the three topics [$F(2,62) = .035$, $p = .966$]. Although the correct rate of “Platypus” was slightly higher than in the pilot study, there was no significant difference in the correct rate between the pilot and the main study ($p = .437$).

3.2.2. Unit of information

All participants stopped listening at the end of the sentences. The number of letters, words, and sentences that participants listened to were respectively calculated and analyzed to compare the effect of three topics on the number of listened to letters, words, and sentences using ANOVA; with the results shown in Table 3. The results revealed significant differences in letters [$F(2,62) = 3.916$, $p = .025$] and words [$F(2,63) = 4.161$, $p = .020$] between three topics. But there

was no significant difference in sentences [$F(2,62) = .557$, $p = .576$].

Scheffe’s Test for multiple comparisons found that the number of letters and words was significantly difference between “Dingo” and “Platypus” [$p_{(letters)} = .045$, $p_{(words)} = .032$]. Although the number of letters and words was higher in “Platypus” than in the other two topics, the number of sentences was rather higher in “Dwarf Pine.”

3.3. Discussion

In Experiment 1, the difference in the accuracy of memory and the number of information units (letters, words, and sentences) between the three topics that participants listened to was analyzed. Through this analysis, it was revealed which type of information units was used and matched for users’ information processing.

The difficulties of the three topics were equal in both studies, and participants tried to remember the information and not to answer randomly. Therefore, the point where the participants stopped listening did not relate to the difficulty of each topic. In addition, it was expected that the correct rate is higher in the main study than in the pilot study, but the results showed the same correct rate in both studies. Although the task was to remember as much information as they can, participants seemed to have chosen the point where they wanted to stop based on their preference. Nevertheless, there was no change in that the point at which they stopped listening was based on the unit of information processing.

The number of sentences that the participants listened to did not differ by three topics, but not the same number of letters or words. These results support that the sentence served as a unit of information. In the case of “Platypus,” the number of sentences was not different, even though the number of letters and words was significantly smaller, so it

Table 2. Correct rate of the pilot study and main study.

Topic	Pilot		Main	
	N	Correct rate (SD)	N	Correct rate (SD)
Dwarf Pine	16	63.13 (13.276)	24	64.41 (13.163)
Dingo	16	67.19 (14.020)	18	65.57 (14.438)
Platypus	16	58.13 (12.500)	23	64.77 (15.090)

Note. Participants in the pilot study listened to all 20 sentences. But they listened to as much information as they wanted in the main study.

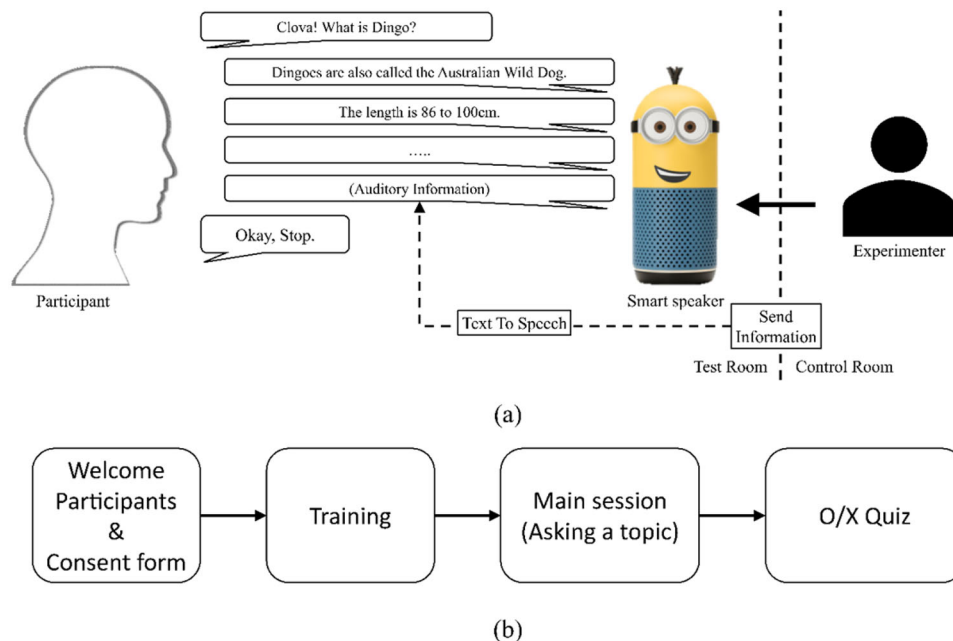


Figure 2. (a) Generic diagram of the experiments and (b) procedure of Experiment 1.

Table 3. ANOVA results of the amount of listened information by sentence, word, and letter.

Unit	Dwarf pine	Dingo	Platypus	<i>F</i>	<i>p</i>
Sentence	13.92 (6.426)	11.94 (5.493)	12.83 (6.088)	0.557	0.576
Word	61.42 (29.635)	53.39 (35.296)	84.78 (44.274)	4.161	0.020*
Letter	206.54 (96.955)	188.39 (107.393)	280.35 (134.457)	3.916	0.025*

* $p < 0.05$.

can be said that participants used the sentence as a criterion for determining memory limitation. This can be described as a human chunking mechanism (Gobet et al., 2001). People create links between elements of information to increase their memory span, which is the chunking mechanism. Likewise, participants may have used chunking, and they would have linked the letters and words in the sentence and processed the sentence at once.

This experiment suggested that the auditory information of VUI systems should consist of sentences, and the sentence is the unit of auditory information. People prefer to use sentences to process and memorize information. However, to provide the auditory information, which is proper for VUI users, it was necessary to investigate the size of the information unit. Accordingly, the aim of Experiment 2 was to explore the size of information units that people can use to process information.

4. Experiment 2

Experiment 1 confirmed that the sentence could serve as a unit of information. In this experiment, modified sets of encyclopedia information were presented auditorily through the smart speaker. The size of unit sentences was limited based on the magic number 7 ± 2 (Miller, 1956) and the information block (number of units per block) was added as an additional condition for the experiment. Analyzing the effects of unit/block on the amount of listened information revealed the proper size of the information unit and the proper amount of information.

4.1. Method

The experiment setting of this experiment was almost identical to Experiment 1 in Figure 1.

4.1.1. Participants

Thirty-seven local university members (19 males and 18 females) participated in this experiment. The overall mean age of the participants was 27.0 years ($SD = 3.59$). All participants reported experience with VUI systems. They were compensated \$15 for about 60 min of their time.

4.1.2. Apparatus and stimuli

Three topics (Platypus, Kangaroo, Kim-Gu) were selected as stimuli. Those were also the information that is not for experts but is unfamiliar to the public. Each topic consisted of 30 sentences which were limited to a maximum of nine words. These sentences were combined into 1, 2, or 3

sentences per block and converted into an information block. Sentences were combined using conjunctions, and the number of letters or words was almost the same as the original information. Therefore, in the 3 units/block condition, 10 information blocks were presented to participants. The example of “Platypus” was shown in Table 4, which was translated from Korean to English.

In this experiment, the same smart speaker as in Experiment 1 was used.

4.1.3. Procedure

The experimental procedure was almost the same as Experiment 1, except that each participant asked all three topics to the smart speaker (in Figure 3). “Dingo” in Experiment 1 was given for training. Following this, three unit/block conditions were presented in random order from the smart speaker in Korean. As in Experiment 1, participants could stop listening by speaking “stop.” After listening to the information on each block condition, participants completed the OX quiz.

4.2. Results

4.2.1. Correct rate

The mean of the correct rate was 71.4% ($SD = 14.43$) slightly higher than Experiment 1 (see Table 5). The effects of topic and unit/block on the correct rate were analyzed, but the results showed that there were no significant main effects [$p_{(topic)} = .180$, $p_{(unit/block)} = .965$] and interaction effects between topic and unit/block [$F(4,102) = 1.305$, $p = .273$]. So, this result means that the difficulties of the three topics were equal and people tried to remember the information.

‘Platypus’ was used in both Experiments 1 and 2. So, its correct rates were analyzed to discover the difference between experiments, but there was no significant difference ($p = 0.243$).

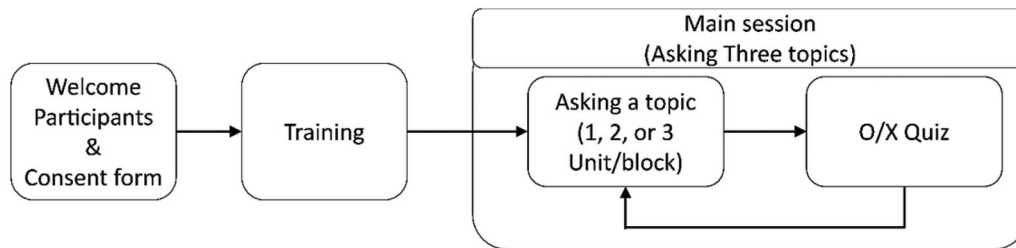
4.2.2. Information unit/block

To determine the size of the information unit, two-way ANOVA was performed to analyze the effects of topic and unit/block on the number of listened information units and blocks. The results were shown in Figure 4.

In terms of the number of information blocks, the ANOVA revealed that there was only a significant main effect of unit/block on the number of information blocks [$F(2,102) = 26.396$, $p < .000$]. But there was no main effect of the topic [$F(2,102) = .394$, $p = .675$] and interaction effect [$F(4,102) = 1.960$, $p = .106$]. Scheffe’s Test for *post-hoc* found that the mean value of information units in “1 unit/block” condition (mean = 15.3, $SD = 8.07$) was significantly

Table 4. Example of auditory information of “Platypus” in Experiment 2.

No	Sentence (translated from Korean to English)	Information unit/block					
		1 Unit/block		2 Units/block		3 Units/block	
		Letters	Words	Letters	Words	Letters	Words
1	Sometimes known as a “Duck-billed” platypus.	11	2	29	8	39	11
2	The platypus is the most primitive of modern mammals.	19	6				
3	The platypus is the mammal that lays eggs.	9	3	19	7		
4	The platypus has a thick body and four short legs.	11	4			42	10
5	The body is 30–45 cm long.	14	3	31	6		
6	The tail is 10–14 cm long.	15	3				
7	Weight varies from 1–1.8 kg	14	3	25	6	35	9
8	Males are larger than females.	10	3				
9	The tail is long and flat.	9	3	17	6		
10	The feet are wide.	8	3			31	9
11	There are five claws on each foot.	12	3	21	6		
12	The webbing is on the feet.	13	3				
13	The webbing of the forefoot is so large that it is out front of the toe.	22	6	32	9	50	15
14	It is folded back when walking.	14	4				
15	The webbing of the hind feet is as small as the length of the toe.	18	6	36	12		
16	The conspicuous spur is located on each inner male’s hind ankle.	18	6			52	17
17	The spurs are linked to the venomous glands and delivers a venom.	16	5	35	11		
18	The snout is wide and flat like a duck.	19	6				
19	The snout is hairless.	9	3	26	9	50	15
20	The snout is covered with sensitive soft skin.	22	6				
21	There are oval nostrils on the front dorsal surface of the snout.	20	6	34	11		
22	The eyes are small and the front of the head.	13	5			43	15
23	The ear without an outer ear lobe is located just behind the eye.	17	6	31	11		
24	They have a short woolly fur coat.	11	4				
25	The dorsal color is taupe.	10	3	32	11	55	18
26	The ventral color is ash gray or tan with a silvery gloss.	23	8				
27	The mouth is wide and has a large cheek pouch.	22	7	41	13		
28	Platypuses lose their teeth as they grow into adults.	18	6			53	19
29	Two pairs of tough keratinized pads are used as teeth.	20	7	36	13		
30	The platypus is most likely to be observed early in the morning or late in the evening.	16	6				

**Figure 3.** Procedure of Experiment 2.**Table 5.** Correct rate depending on topic and unit/block.

		Correct rate (SD)
Topic	Platypus	69.6% (15.59)
	Kangaroo	75.5% (13.53)
	Kim Gu	69.0% (13.55)
Unit/block	1 unit	72.0% (15.69)
	2 units	70.5% (13.42)
	3 units	71.6% (14.44)
	Total	71.4% (14.43)

different from the other two conditions (2 unit/block: mean = 8.4, $SD = 3.65$; 3 unit/block: mean = 6.6, $SD = 2.77$; all $ps < .000$).

On the contrary, there were no significant effects of unit/block [$F(2,102) = 3.088$, $p = 0.050$], topic [$F(2,102) = .632$, $p = .533$], and the interaction between them [$F(4,102) = 2.413$, $p = .054$] on the number of information units. The average number of information units was 17.3 sentences ($SD = 8.05$).

4.3. Discussion

Experiment 2 extended the findings of Experiment 1 in the size of the information unit. An appropriate size of the information unit was investigated in this experiment by constructing the information blocks based on the sentence, which serves as the information unit revealed in Experiment 1.

Correct rates were not significantly different between topics in Experiment 2. Based on the results of “Platypus,” it also seems that there would be no significant difference in difficulties between experiments 1 and 2. Additionally, both experiments did not show a high rate of correct answers for each participant or topic, which means that the participants did not have prior knowledge about the subject. Thus, the results showed that participants had decided to stop listening to information regardless of the difficulties of the topics.

In terms of information unit/block, there were two findings discovered by this experiment. First, the size of the

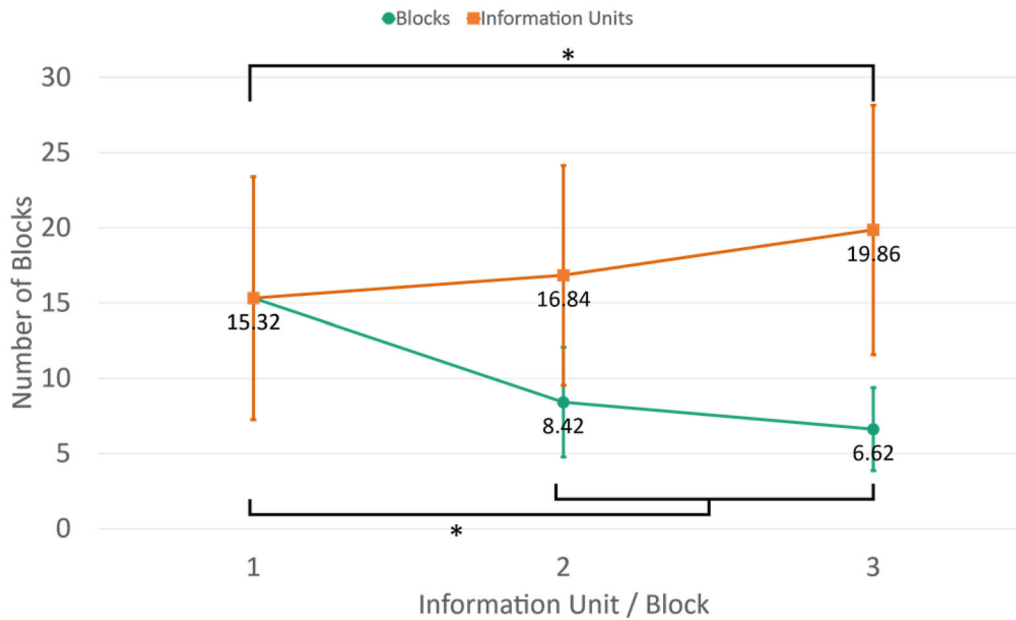


Figure 4. The number of listened information depending on unit/block.

information unit is not allowed to be extended in VUI. As the number of information units per block increased, the number of information blocks listened to by the participants decreased. Although three units per block condition showed slightly higher than others, there were also no significant differences in the number of listened information units. These results mean that the extension of the size of the information unit had no benefit for participants' information processing. When people listen to information, they usually relied on short-term memory for the auditory information, which is linear and ephemeral (Jung et al., 2020). This can be the reason there are no benefits of extension of the information unit in this experiment. Although the researchers converted information units to information blocks, the information units were presented to participants in order and they could still process the information units separately.

Second, participants wanted to listen to about 17 sentences on average. The mean correct rate was 71.4%, so participants were expected to be able to remember 12–13 sentences at a time. Sentences were refined to a certain size, and participants listened to more sentences in Experiment 2 than in Experiment 1. Nevertheless, the correct rates remained the same or higher than those in Experiment 1, and consequentially participants remembered approximately the same amount of information as in the pilot study (about 12 sentences: listening 20 sentences and 62.8% correct rate). These results indicated that refined information gave less burden to participants' information processing and that allowed them to determine their memory capacity precisely and to remember the information accurately. This also supported the previous studies that the length or size of information places a burden on the user's working memory (Baddeley et al., 1975; Montgomery, 2004).

5. Conclusion

While existing studies focused on technical aspects or input methods, the present study explored the information recognized by users and suggests the appropriate unit and size of information for VUI system. It was determined that the sentence limited to nine words can serve as a unit of auditory information in VUI. Furthermore, users could remember more information when it was refined. This suggests the importance of user-centered auditory information design to researchers or developers in the field of VUI. Based on these discussions, this study proposes three design guidelines to decrease the cognitive load of VUI from the auditory perspective.

1. Auditory information should be in sentence form.
2. The sentence consists of up to nine words.
3. A maximum of 17 sentences can be presented, but users can only remember 70% (about 12 sentences)

These findings are expected to help design information when delivering safety guidelines or device instructions that require users to remember as much information as possible. In addition, it is expected to help in education and training using smart speakers or VUI by predicting the amount that can be provided to users or that users will remember.

To focus on the unit and size of information, the effects of various characteristics of speech were not considered in this study. For example, pitch or speed of speech could be cues for inferring others' personalities of others and those could affect the expertness of messages (Kim et al., 2021). In addition, given that usual video services have a speed control function, users may be able to process the information quickly even if the speaking rate increases. For this reason,

the effect of various characteristics of speech on the size of the information unit should be investigated in future studies.

Additionally, encyclopedia topics were used for providing a lot of information at once in this study. However, such information might be felt familiar depending on the participant's education, job, or interest, and accordingly, differences in the level of knowledge might occur. To minimize the influence of information in specific areas, further studies should be conducted using information that people need daily, such as news, traffic, and weather, that changes in real-time, and is not limited to a specific area (Park et al., 2021).




Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partly supported by the Institute for Information & Communications Technology Promotion funded by the Korean government (MSIP) (R0124-16-0002, Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly).

ORCID

Min Chul Cha  <http://orcid.org/0000-0001-9301-4281>
 Hyo Chang Kim  <http://orcid.org/0000-0002-4279-6620>
 Yong Gu Ji  <http://orcid.org/0000-0002-0697-2164>

References

- Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019, May 2). Guidelines for human-AI interaction. In *Conference on Human Factors in Computing Systems - Proceedings*. <https://doi.org/10.1145/3290605.3300233>
- Asabuki, T., Hiratani, N., & Fukai, T. (2018). Interactive reservoir computing for chunking information streams. *PLOS Computational Biology*, 14(10), e1006400. <https://doi.org/10.1371/journal.pcbi.1006400>
- Aylett, M. P., Kristensson, P. O., Whittaker, S., & Vazquez-Alvarez, Y. (2014). None of a CHInd: Relationship counselling for HCI and speech technology. In *CHI 2014* (pp. 749–760). <https://doi.org/10.1145/2559206.2578868>
- Baddeley, A. D. (2000). Short-term and working memory. In *The Oxford handbook of memory* (Vol. 4, pp. 77–92). Oxford University Press.
- Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Elsevier. [https://doi.org/10.1016/S0079-7421\(08\)60452-1](https://doi.org/10.1016/S0079-7421(08)60452-1)
- Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14(6), 575–589. [https://doi.org/10.1016/S0022-5371\(75\)80045-4](https://doi.org/10.1016/S0022-5371(75)80045-4)
- Bickmore, T. W., Pfeifer, L. M., & Jack, B. W. (2009). Taking the time to care: Empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1265–1274). https://doi.org/10.1007/978-3-642-12770-0_8
- Cohen, M. H., Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Addison-Wesley Professional.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1), 87–114. <https://doi.org/10.1017/S0140525X01003922>
- Dirlam, D. K. (1972). Most efficient chunk sizes. *Cognitive Psychology*, 3(2), 355–359. [https://doi.org/10.1016/0010-0285\(72\)90012-6](https://doi.org/10.1016/0010-0285(72)90012-6)
- Gilchrist, A. L., Cowan, N., & Naveh-Benjamin, M. (2009). Investigating the childhood development of working memory using sentences: New evidence for the growth of chunk capacity. *Journal of Experimental Child Psychology*, 104(2), 252–265. <https://doi.org/10.1016/j.jecp.2009.05.006>
- Glanzer, M., & Razel, M. (1974). The size of the unit in short-term storage. *Journal of Verbal Learning and Verbal Behavior*, 13(1), 114–131. [https://doi.org/10.1016/S0022-5371\(74\)80036-8](https://doi.org/10.1016/S0022-5371(74)80036-8)
- Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6), 236–243. [https://doi.org/10.1016/S1364-6613\(00\)01662-4](https://doi.org/10.1016/S1364-6613(00)01662-4)
- Hauptmann, A. G., & Rudnick, A. I. (1990). A comparison of speech and typed input. In *HLT '90: Proceedings of the Workshop on Speech and Natural Language* (pp. 219–224). <https://doi.org/10.3115/116580.116652>
- Howell, M., Love, S., & Turner, M. (2006). Visualisation improves the usability of voice-operated mobile phone services. *International Journal of Human-Computer Studies*, 64(8), 754–769. <https://doi.org/10.1016/j.ijhcs.2006.03.002>
- Hua, Z., & Ng, W. L. (2010). Speech recognition interface design for in-vehicle system. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '10, AutomotiveUI* (p. 29). <https://doi.org/10.1145/1969773.1969780>
- Jung, J., Lee, S., Hong, J., Youn, E., & Lee, G. (2020). Voice + Tactile: Augmenting in-vehicle voice user interface with tactile touchpad interaction. In *Conference on Human Factors in Computing Systems - Proceedings* (pp. 1–12). <https://doi.org/10.1145/3313831.3376863>
- Kim, H. C., Cha, M. C., & Ji, Y. G. (2021). The impact of an agent's voice in psychological counseling: Session evaluation and counselor rating. *Applied Sciences*, 11(7), 2893. <https://doi.org/10.3390/app11072893>
- Kim, J., Jeong, M., & Lee, S. C. (2019). “Why did this voice agent not understand me?”: Error recovery strategy for in-vehicle voice user interface. In *Adjunct Proceedings - 11th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2019* (pp. 146–150). <https://doi.org/10.1145/3349263.3351513>
- Kwon, O. W., & Park, J. (2003). Korean large vocabulary continuous speech recognition with morpheme-based recognition units. *Speech Communication*, 39(3–4), 287–300. [https://doi.org/10.1016/S0167-6393\(02\)00031-6](https://doi.org/10.1016/S0167-6393(02)00031-6)
- Lee, S. C., & Ji, Y. G. (2019). Complexity of in-vehicle controllers and their effect on task performance. *International Journal of Human-Computer Interaction*, 35(1), 65–74. <https://doi.org/10.1080/10447318.2018.1428263>
- Lo, V. E. W., & Green, P. A. (2013). Development and evaluation of automotive speech interfaces: Useful information from the human factors and the related literature. *International Journal of Vehicular Technology*, 2013, 1–13. <https://doi.org/10.1155/2013/924170>
- López, G., Quesada, L., & Guerrero, L. A. (2018). Alexa vs. Siri vs. Cortana vs. Google Assistant: A comparison of speech-based natural user interfaces. *Advances in Intelligent Systems and Computing*, 592, 241–250. https://doi.org/10.1007/978-3-319-60366-7_23
- Mane, A., Boyce, S., Karis, D., & Yankelovich, N. (1996). Designing the user interface for speech recognition applications. In *CHI '96: Conference Companion on Human Factors in Computing Systems* (Vol. 28, p. 431). <https://doi.org/10.1145/257089.257431>
- Miller, G. (1956). Human memory and the storage of information. *IEEE Transactions on Information Theory*, 2(3), 129–137. <https://doi.org/10.1109/TIT.1956.1056815>
- Montgomery, J. W. (2004). Sentence comprehension in children with specific language impairment: Effects of input rate and phonological working memory. *International Journal of Language &*

- Communication Disorders*, 39(1), 115–133. <https://doi.org/10.1080/13682820310001616985>
- Park, J., Choi, H., & Jung, Y. (2021). Users' cognitive and affective response to the risk to privacy from a smart speaker. *International Journal of Human-Computer Interaction*, 37(8), 759–771. <https://doi.org/10.1080/10447318.2020.1841422>
- Pearl, C. (2016). *Designing voice user interfaces: Principles of conversational experiences*. O'Reilly Media, Inc. https://books.google.co.kr/books?hl=ko&lr=&id=MmnEDQAAQBAJ&oi=fnd&pg=PR11&dq=VUI+advantage+multitasking&ots=HNb-0ubBhd&sig=GkFodoix6KBzwMh-UC_yDYZaBhU#v=onepage&q=VUIadvantagemultitasking&f=false
- Peissner, M., & Doebler, V. (2011). Can voice interaction help reducing the level of distraction and prevent accidents? *Whitepaper*, May 24.
- Perlman, D., Samost, A., Domel, A. G., Mehler, B., Dobres, J., & Reimer, B. (2019). The relative impact of smartwatch and smartphone use while driving on workload, attention, and driving performance. *Applied Ergonomics*, 75(September 2018), 8–16. <https://doi.org/10.1016/j.apergo.2018.09.001>
- Perugini, S., Anderson, T. J., & Moroney, W. F. (2007). A study of out-of-turn interaction in menu-based, IVR, voicemail systems. In *Proceedings of ACM CHI 2007: Conference on Human Factors in Computing Systems* (pp. 961–970).
- Ranney, T. A., Harbluk, J. L., & Noy, Y. I. (2005). Effects of voice technology on test track driving performance: Implications for driver distraction. *Human Factors*, 47(2), 439–454. <https://doi.org/10.1518/0018720054679515>
- Saariluomaand, P., & Jokinen, J. P. P. (2014). Emotional dimensions of user experience: A user psychological analysis. *International Journal of Human-Computer Interaction*, 30(4), 303–320. <https://doi.org/10.1080/10447318.2013.858460>
- Sawan, P. B. D., Gopy, K., Hurry, G., & Gopaul, T. T. (2013). A study on smart home control system through speech. *International Journal of Computer Applications*, 69(19), 30–39. <https://doi.org/10.5120/12080-8244>
- Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM*, 43(9), 63–65. <https://doi.org/10.1145/348941.348990>
- Strayer, D. L., Cooper, J. M., McCarty, M. M., Getty, D. J., Wheatley, C. L., Motzkus, C. J., Goethe, R. M., Biondi, F., & Horrey, W. J. (2019). Visual and cognitive demands of CarPlay, Android Auto, and Five Native Infotainment Systems. *Human Factors*, 61(8), 1371–1386. <https://doi.org/10.1177/0018720819836575>
- Voicebot.ai. (2020a, January). *In-car voice assistant consumer adoption report January 2020*. Voicebot.ai. https://voicebot.ai/wp-content/uploads/2020/02/in_car_voice_assistant_consumer_adoption_report_2020_voicebot.pdf
- Voicebot.ai. (2020b, April). *Smart speaker consumer adoption report executive summary April 2020*. <https://www.voice2shop.com/wp-content/uploads/2020/11/executive-summary-smart-speaker-consumer-adoption-report-2020.pdf>
- Watanabe, M., Okano, A., Asano, Y., & Ogawa, K. (2007). VoiceBlog: Universally designed voice browser. *International Journal of Human-Computer Interaction*, 23(1–2), 95–113. <https://doi.org/10.1080/10447310701362975>
- Wickelgren, W. A. (1964). Size of rehearsal group and short-term memory. *Journal of Experimental Psychology*, 68(4), 413–419. <https://doi.org/10.1037/h0043584>
- Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2), 159–177. <https://doi.org/10.1080/14639220210123806>
- Winsum, W. V., Martens, M., & Herland, L. (1999). The effects of speech versus tactile driver support messages on workload, driver behaviour and user acceptance. In *TNO Human Factors*. <https://doi.org/10.13140/RG.2.1.1776.1041>
- Yager, C. (2013). *An evaluation of the effectiveness of voice- to-text programs at reducing incidences of distracted driving* (Report SWUTC/13/600451-00011-1). Texas A&M Transportation Institute. <http://swutc.tamu.edu/publications/technicalreports/600451-00011-1.pdf>
- Yankelovich, N., Levow, G. A., & Marx, M. (1995). Designing speechActs: Issues in speech user interfaces. In *Conference on Human Factors in Computing Systems – Proceedings* (Vol. 1, pp. 369–376).

About the authors

Min Chul Cha is a PhD candidate in the Department of Information and Industrial Engineering at Yonsei University, Seoul, Korea. His research interests include voice user interfaces and usability/UX in smart devices.

Hyo Chang Kim is a research scientist at the Stanford Center at Incheon Global Campus (SCIGC). He received his PhD in Industrial Engineering from Yonsei University. His research interests include usability/UX in HRI and autonomous vehicles.

Yong Gu Ji is a professor in the Department of Industrial Engineering at Yonsei University, where he directs the Interaction Design Laboratory. He received his PhD in Human Factors/HCI from Purdue University. His research interests include usability/UX in smart devices and self-driving vehicles.