International Journal of Engineering & Technology, *8 (1.4) (2019) 471-477*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Robot Auditory Interface Design Factor Attributes and the Level Values Based on UX using TTS System

Seung Eun Chung[1], Han Young Ryoo[2]*

[1,2] Dept. of Content Convergence, Ewha Womans University,
52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760 Republic of Korea.
**Corresponding author E-mail: hyryoo@ewha.ac.kr*

Abstract

The purpose of this study is to suggest the attributes of the design factors and the level values for each attribute for the service robot's auditory interface design. To do so, the main design factors of the auditory interface were organized through the research on the existing literature, and then the attributes and the level values of the auditory interface design factors that can be designed using the SSML from the TTS system were organized. A survey was conducted to verify whether the organized values have a significance meaning in the aspect of the user experience, and the level values which had differences in all the functional/service experience, the interaction experience, and the emotional experience have been organized. From the results, 'Prototype of the sound effect', 'Gender', 'Pitch', 'Pitch range', and 'Rate' were verified to be useable as the attributes of the auditory interface design factors to measure the overall user experience with the robots. Two out of the 15 investigated subordinate level values had been deleted and integrated so that a total of 13 level values have been suggested in the end. This research has a significant meaning in a way that it offers suggestions to the designers of the robots that provide auditory interface through the TTS system the usable design factors for the user experience-centered designs.

Keywords: *Social Service Robot, Auditory Interface, Design Factors and values, User Experience*

1. Introduction

The auditory interface is the interface where the communication between the human and the machine takes place through the medium of sound or voice. This interface enables to relay information whether or not the listener is fully paying attention so it is very useful for the robots that can move around. Also, since it allows communicating while working on something else, it is becoming more important in the service context.

The existing researches regarding the auditory interface mainly dealt with the technical subjects like the voice recognition or the voice synthesis [1][2] and focused on the aspect of the usability resolving the problems when using the auditory interface and making it more convenient to use[3][4]. However, as the users recently started to have a tendency to consider the intelligent products or robots that use the auditory interface as living organisms [5], the impression and the experiences that the robot voice generates are starting to become important.

Generally speaking, a voice is used as a measurement to evaluate the impression or image. Humans can delicately differentiate the features of a sound such as volume, height, length, and timbre with ears [6] and through this differentiation process, they can understand the characteristic information like emotions, health, aging, and etc. [7] When communicating, only 7% out of all the messages is delivered by the contents of the language, 55% is delivered by the visual elements such as the facial expression, the posture, the gesture or the look, and the rest 38% is delivered by the auditory elements such as the voice or the tone of the voice [8]; after all, the voice occupies a significant portion of a person's image formation.

Thanks to the recent improvement of the voice recognition rate and the development of the voice player devices, the voice interface is actively being applied to the robots or various types of the intelligent products. As the accuracy of the TTS (Text to Speech) technique, one of the voice expression techniques, is being gradually improved and the cloud services that provide the API are expanding, the opportunities of the voice interface to be used in the service robot operated by voice are increasing.

Yet, the researches on the design factors that express the voice to experience the robot's impression or characteristics are not conceptually systematized and there were hardly any studies specifically focused on the aspect of the design factors that could be used in the TTS services. The TTS services offer various expressions by allowing the control of the detailed characteristics of the voice; therefore, it is necessary to systematize the utilizable factors based on the understanding of the meaning of the attributes suggested within the services.

Thus, this research is to find out the design factors that are mainly discussed in the literature as the robot auditory interface factors, and to derive the robot auditory interface design factors that can be practically used in the TTS services as well as their detailed attributes. A user survey was conducted to systematize the auditory design factors that can be utilized in the user experience design by suggesting the attributes and the detailed level values of the factors that show differences in the aspect of the user experiences.

2. Systematization of the Robot Auditory Interface Design Factors

2.1. Auditory Interface Design Literature Research

In this study, the literature research was conducted to understand the factors that are discussed to be used in designing the interface where the robot communicates through sounds.

Arons & Mynatt(1994) had divided the auditory factors that can be perceived through the sound interface roughly into the speech and non-speech factors. The speech factors are effective when delivering detailed information and the non-speech factors are effective when delivering emergency information due to its short play time [9]. In order to make use of these characteristics effectively, many of the robot's voice interfaces offer both the speech and non-speech sounds.

Garzonis et al. (2009) said that the non-speech sound can be divided into the auditory icons and the earcon. The auditory icon is the sound effect that combines the sounds that can be heard in everyday life with the information [10]. In the promotional video of the robot Jibo, it makes the clicking sound when taking a picture; this is an example of the auditory icon. On the other hand, the earcon is an abstract sound which is the sound effect produced by the structured combinations of the various heights of the sound in order to relay the information about the interaction with the computer [11]. Because the auditory icon can connect the meaning of the information, it is effective when it comes to learning and memory, whereas the earcon does not have any meaning in the indicated information, so the user should learn and remember it. But, some say that the earcon is suitable to keep the system consistent [12].

Kwak et al. (2012) said that the anthropomorphic sound is important since the user sees the robot as a living organism and attempts to socially interact and divided the non-speech sound into the anthropomorphic sound and the non-anthropomorphic sound. The anthropomorphic sound is the sound that makes the product feel like a living organism by applying the anthropomorphic factors to the sound of a product. When the robot 'Ijini' is petted, it makes the sound of a sigh or something like 'mmm' just like the sound that a person makes when feeling good; this is an example of an anthropomorphic sound. They also mentioned that the anthropomorphic non-speech sound is more effective and similar to the speech sound when it comes to empathizing with the robot than the non-anthropomorphic sound [5].

Considering that the voice interface with the speech sound conveys information, the factors that affect this interface are importantly being researched. Chae et al. (2007) said that depending on the own physical characteristics of a voice, the cognitive efficiency of the delivered information and the emotional satisfaction can differ and they measured the preference in the product's voice interface with the major physical characteristics such as the gender (female and male), the height of the voice (tone: common, high), and the tone of the voice (intonation: flat, dynamic) [13]. Jee et al. (2010) defined the factors with a focus on the relationship between the perceived feelings and the musical features of a voice. They analyzed the sounds of the robot in the movies and came into a conclusion that the intonation, the pitch, and the timbre are appropriate to express the intention and emotions [14].

The research on the voice analysis factors that affect the listeners have been developed in more detail with the voice analysis technology. Choi, Cho, & Jeong (2016) said that the generally used voice analysis factors can be divided into height, variance, and intensity of the tone as well as jitter, shimmer, NHR(Noise to Harmony Ratio), and speed that measure the tone and their effectiveness have been verified [13]. These factors can also be expressed by the prosody factors, and Schmitz, Krüger, Schmidt (2007) stated that the prosody factors have a big impact on the perceived characteristics and can be used partially in order to get the desired impression. The tested prosody factors are the pitch range, pitch level, tempo and intensity. They all can characterize the products that use the auditory interface and can be used as the important expression factors as well [16].

These factors can create different impressions to the users by the combination of the detailed levels. The pitch reveals the height of a voice. Especially, if a male has a low level of pitch, it tends to be interpreted to give a sense of stability and trust [17]. When the pitch range is rather flat, it sounds logical and stable [17], but very active when the range is dynamic [18]. The intensity is the strength in a voice; a high intensity gives a very strong feeling but a low intensity gives a very soft feeling [19]. Moreover, the timbre is closely related to the public confidence. The richness of the tone increases the public confidence of the speaker and it can affect the listener in a good way [15]. For the speaking speed, it feels intellectual and objective at a fast speed [20] whereas it feels prudent and stable at a slow speed [15].

2.2. Understanding of TTS System and SSML

The Text-To-Speech (TTS) system is a system in which a certain text is input and being transferred to generate its corresponding voice as the output and it is also called the voice synthesis system [21]. This technique is offered in the service where the voice application can be developed for the robots or user products that provide a voice interface. This service is offered at a low cost and invoiced and used by various business operators of the intelligent products and robots to provide a voice interface.

These services support the Speech Synthesis Markup Language (SSML – Standardized language for voice synthesis based on XML settled by W3C) for the voice synthesis expression. This SSML offers markups such as the sentence structure analysis, text preprocessing, pronunciation method statement, prosody control, and miscellaneous audio file output. The process of the SSML documents is as follows: structure analysis, text preprocessing, pronunciation conversion, prosody analysis, and voice waveform creation. Especially, it can produce characteristic synthetic voices by adjusting the level values of the SSML's prosody elements such as pitch, rate, volume, contour, range, and duration [21]. In other words, the emotions and strength can be synthesized as desired through the voice expression mark-up language called 'SSML'. These elements can be only utilized based on the terms of use set by W3C [23]. The factors that can be used for the style of the voice expression are the 'Prosody and Style' and the detailed attributes and the level values are organized as in Table 1 below.

Table 1: Guideline for the SSML Prosody and Style

Division	Attribute	Values
	Gender	The gender (Ex. "Male", "Female", "Neutral")
	Age	The order of age (Ex. 0, 1, 2, etc.)
	Variance	The variance of the other voice characteristics (Ex. 0, 1, 2, etc.)
	Name	The names for the voice by process (Ex. A name can be made for a specific effect)
	Languages	The list of the languages (Ex. English with a Portugal accent and English with a Japanese accent and etc. can be distinguished)

Emphasis	Level	Strength of stress ("Strong", "Moderate", "None")
Break	Strength	The strength of the prosody violation of the voice output. The break boundary strength of the output is decided. ("None" , "X-weak" , "Weak" , Medium", "Strong" or "X-strong")
	Time	The pause time (Ex. "250ms", "3s")
Prosody	Pitch	The increase or decrease of the output pitch. ("Hz" after a number, Relative change or "X-low", "Low" , "Medium" , "High" , "X-high")
	Contour	The pitch contour is defined as a set of white space-separated targets at specified time positions in the speech output
	Range	The highness or lowness of a pitch of the included text (Variability). ("Hz" after a number, Relative change or "X-low" , "Low" , "Medium" , "High" , "X-high")
	Rate	The change of the speaking speed. (Non-negative numbers but ratio or "X-slow" , "Slow" , "Medium" , "Fast" , "X-fast")
	Duration	The time that takes to read the included text. (Ex. "250ms", "3s")
	Volume	The loudness and softness of a sound. (Ex. "Silence" , "X-soft" , "Soft" , "Medium" , "Loud" , "X-loud" or "+/-" and "dB")

However, the TTS services only provide a few complete set of voices that define the basic voice attributes rather than having all SSML tags open to support. "Amazon Polly", one of the representative TTS cloud services, supports 24 languages and 47 types of specific voices. Even within the same language, it is provided with different region options because of their pronunciation differences. For example, it separated the British English from the American English and these are again distinguished by gender or timbre and the voices are produced accordingly and given a name for each voice. For the American English, a total of five female voice characters with the names of Salli, Kimberly, Kendra, Joanna, and Ivy, and a total of three male voice characters called Matthew, Justin and Joey are offered. The designer or the developer can choose a name of a character by language and gender, and then modify the detailed attributes of the voice expression through the tags of emphasis, break, and prosody. It is necessary to make good use of the prosody factors because it can affect the expressed impression or the user experience depending on the choices of the detailed level value for each attribute. In Amazon Polly, only the selected attribute expressions of pitch, rate, and volume are supported as part of the SSML-prosody attributes.

Another TTS service called "IBM Bluemix" supports nine languages and provides 13 different types of voices. IBM presents 'Expressive SSML' and 'Voice transformation SSML' that they defined on their own as a set of tags for a style along with some factors of the SSML, and guide the users to designate the desired one. From the SSML-Prosody attributes, pitch, rate and volume are supported. In addition, the TTS simulator from Action on Google provides a single voice per 22 languages and the pitch, rate and volume from the SSML-Prosody attributes are supported.

2.3. Auditory Interface Design Factors and Attributes and Level Value Organization

Next, based on the factors and the attributes discussed through the literature analysis, the attribute elements and the subordinate level values that can be currently used in the TTS services are organized.

The non-speech sound factors can be divided into the anthropomorphic and non-anthropomorphic sounds; and again, the non-anthropomorphic sound can be sub-divided into the auditory icon, which can be heard in daily life, and the earcon, which is abstractly structured. Since these factors are related to the prototype that makes sounds or the prototype from which the sounds are being made, the attributes can be defined as the prototype of the sound effect.

The speech sound factors affect the delivery of the impression or the emotions of the robot, the disputed expression attributes are the genders, pitch, pitch range (intonation), timbre, intensity, tempo of the voice. Out of these attributes, the ones that can be controlled by the robot service provider as the SSML standard tag in the TTS system are the gender, pitch, pitch range, volume (intensity), and rate(tempo). The volume, however, is also removed due to the fact that it can be controlled by the user depending on the surrounding environment. The subordinate level values are organized using the level values that are expressed in the SSML. The above details are organized as in Table 2 below.

Table 2: Attributes and the Level Values of the Auditory Interface Factors

Factor	Attribute	No	Values
			Definition of the Level Value
Non-speech sound	Prototype of the sound effect	1_1	Anthropomorphic sound
			The non-verbal sound made by imitating a person or an animal
		1_2	Auditory icon
			The non-verbal sound of objects
		1_3	Earcon
			The mechanical sound that is not by imitating anything else
Speech sound	Gender of voice	2_1	Female Voice
			The sound is of a female voice
		2_2	Male Voice
			The sound is of a male voice
		2_3	Robot-like Voice
			The sound is mechanical and artificial like a robot and is hard to tell whether it is of a male or female voice
	Pitch of voice	3_1	Low of pitch
			The pitch of the voice is low
		3_2	Medium of pitch
			The pitch of the voice is average
		3_3	High of pitch
			The pitch of the voice is high
	Pitch range of voice	4_1	Flat of pitch range
			The rage of the pitch is relatively smooth
		4_2	Medium of pitch range

			The range of the pitch is relatively average
		4_3	Dynamic of pitch range
			The range of the pitch is relatively big
	Rate of voice	5_1	Slow of rate
			The voice rate is slow
		5_2	Medium of rate
			The voice rate is average
		5_3	Fast of rate
			The voice rate is fast

3. User Experience Assessment Survey on Robot Interface Design Factor

The user experience was to be measured against the level values for attributes of the robot's auditory interface design factors that have been organized in the previous step.

For the survey, the plural sample sounds were produced for the 15 level values under the five classifications by using the TTS services by 'Amazon' and 'IBM'. Within the verbal category, the platform standby phrases such as "May I Help you?" and "Please let me know if there's anything else I can do for you." and the feedback phrases such as "Yes, I will carry out the order from now on." were modified as SSML. In order to avoid bias in the difference between the male and female voices, multiple male and female voice samples were provided. Meanwhile, the sample sounds for the non-verbal category were chosen and collected from the 'Sound Library' which provides sound effects offered by 'Amazon' and 'Google'.

A website survey system was created which had a sound discrimination test on the very first page to verify that the sounds were played and heard prior to the survey. The questions were given in the consecutive order after all the sound samples were played. The user experiences with general characteristics were categorized as the functional/service experience, interactional experience, and emotional experience and the expected experiences were measured on a 7-point Likert scale. A total of 224 men and women in their 20s to 40s, who had either direct or indirect experiences with the service robot participated in this survey.

The differences of the user experiences of each level value per attribute were verified by the One-Way Analysis of Variance and the differences between the levels were verified by the Scheff verification. The Scheff verification is a post-hoc analysis which verifies the existence of differences in each attribute by F verification of One-Way Analysis of Variance to find out which group shows the differences through the multiple comparisons. If the significance probability derived by the Scheff verification is less than the significance level of 0.05, it is considered that it does have a difference between the levels. On the contrary, if there is no difference found the Scheff verification, the same levels are grouped together and given the same alphabet; if not, they are given different alphabets.

For the case that each level value per attribute does now show any difference in three experiences of functional/service experience, interactive experience, and emotional experience and have the same level values to be bound into the same group, they are integrated into one level value. Also, if all the level values are integrated and it only has one level value after all under one subdivision, this attribute is decided to be removed.

3.1. User Experience Evaluation on the Prototype of the Sound Effect

The average values of the voices of 1-1. Anthropomorphic sound, 1-2. Auditory icon, 1-3. Earcon, classified by the prototype of the sound effect, were compared. There was a statistically significant difference (p.value: 0.001<) in the difference of the prototype of the sound effect against the A) functional/service experience and the B) interaction experience. According to the Scheff verification, the 'Anthropomorphic sound' and the 'Auditory icon' were confirmed to have the same level (group a) and the 'Earcon' was verified to have a different level (group b). Although the difference of the prototype of the sound effect against the C) emotional experience also had a statistically significant difference (p.value: 0.001<), each level value was at all different levels. In other words, the classified level values can be grouped together in the functional/service experience and the interactional experience. However, in order to evaluate the emotional experience as well, the existing level value classification can have a significant difference.

Table 3: Level Values of the Prototype of Sound Effect Analysis Result

UX	No.	Mean	Scheffe				F	P
			Group	No.	MD	P		
F	1_1	4.272	a	*1_2*	-.038	.951	15.15	.000
				1_3	.549	.000		
	1_2	4.310	a	*1_1*	.038	.951		
				1_3	.587	.000		
	1_3	3.723	b	*1_1*	-.549	.000		
				1_2	-.587	.000		
I	1_1	4.241	a	*1_2*	-.080	.778	14.53	.000
				1_3	.484	.000		
	1_2	4.321	a	*1_1*	.080	.778		
				1_3	.565	.000		
	1_3	3.757	b	*1_1*	-.484	.000		
				1_2	-.565	.000		
E	1_1	4.134	a	*1_2*	.359	.015	23.52	.000
				1_3	.846	.000		
	1_2	3.775	b	*1_1*	-.360	.015		
				1_3	.487	.000		
	1_3	3.288	c	*1_1*	-.846	.000		
				1_2	-.487	.000		

* F: Functional / I : Interaction / E : Emotional

3.2. User Experience Evaluation on Gender of Voice

The average values of the voices of 1-1. Female voice, 1-2.Male voice, and 1-3. Robot-like voice, classified by gender, were compared. The difference of the gender against the A)functional/service experience, B)interactional experience and the C)emotional experience had a statistically significant difference(p.value : 0.001<). According to the Scheff verification, the ‘Female voice’ and the ‘Male voice’ were confirmed to have the same level (group a), and the ‘Robot-like voice’ was verified to have a different level (group b). Therefore, the classification of the level values by gender has been adjusted from ‘female voice, male voice, robot-like voice’ to ‘human voice and robot-like voice’.

Table 4: Level Values of Gender of Voice Analysis Result

UX	No.	Mean	Scheffe				F	P
			Group	No.	MD	P		
F	2_1	4.944	a	*2_2*	.214	.166	66.74	.000
				2_3	1.221	.000		
	2_2	4.730	a	*2_1*	-.214	.166		
				2_3	1.006	.000		
	2_3	3.723	b	*2_1*	-1.221	.000		
				2_2	-1.007	.000		
I	2_1	4.857	a	*2_2*	.203	.778	53.73	.000
				2_3	1.100	.000		
	2_2	4.654	a	*2_1*	-.203	.778		
				2_3	.897	.000		
	2_3	3.757	b	*2_1*	-1.100	.000		
				2_2	-.897	.000		
E	2_1	4.549	a	*2_2*	.268	.090	59.58	.000
				2_3	1.261	.000		
	2_2	4.281	a	*2_1*	-.268	.090		
				2_3	.993	.000		
	2_3	3.288	b	*2_1*	-1.261	.000		
				2_2	-.993	.000		

* F: Functional / I : Interaction / E : Emotional

3.3. User Experience Evaluation on Pitch of Voice

The average values of the voices of 1-1. Low of pitch, 1-2. Medium of pitch, and 1-3. High of pitch were compared. There was a statistically significant difference(p.value: 0.001<) in the different levels of the pitch against the A)functional/service experience and the B)interactional experience. According to the Scheff verification, the ‘Low’ and the ‘High’ were confirmed to have the same level (group a), and the ‘Medium’ was verified to have a different level (group b). The difference of the pitch against the C) emotional experience had a statistically significant difference(p.value: 0.001<) and the levels for ‘low’(group a), ‘medium’(group c), and ‘high’(group b) were all different. In other words, the classified level values can be grouped together in the functional/service experience and the interactional experience. However, in order to evaluate the emotional experience as well, the existing level value classification was determined to be kept.

Table 5: Level Values of Pitch of Voice Analysis Result

UX	No.	Mean	Scheffe				F	P
			Group	No.	MD	P		
F	3_1	3.623	a	*3_2*	-.743	.000	19.20	.000
				3_3	-.234	.162		
	3_2	4.366	b	*3_1*	.743	.000		
				3_3	.509	.000		
	3_3	3.857	a	*3_1*	.234	.162		
				3_2	-.509	.000		
I	3_1	3.580	a	*3_2*	-.815	.000	23.38	.000
				3_3	-.257	.109		
	3_2	4.395	b	*3_1*	.815	.000		
				3_3	.558	.000		
	3_3	3.837	a	*3_1*	.257	.109		
				3_2	-.558	.000		
E	3_1	3.085	a	*3_2*	-.984	.000	30.74	.000
				3_3	-.395	.008		
	3_2	4.070	b	*3_1*	.984	.000		
				3_3	.589	.000		
	3_3	3.480	c	*3_1*	.395	.008		
				3_2	-.589	.000		

* F: Functional / I : Interaction / E : Emotional

3.4. User Experience Evaluation on Pitch Range of Voice

The average values of the voices of 1-1. Flat of pitch range, 1-2. Medium of pitch range, and 1-3. Dynamic of pitch range, classified by the pitch range, were compared. A statistically significant different(p.value : 0.001<) was verified in the different pitch range against the A)functional/service experience, the B)interactional experience and the C)emotional experience. According to the Scheff verification, the ‘Medium’ and the ‘High’ were confirmed to have the same level (group b), and the ‘Flat’ was verified to have a different level (group a).

Thus, the classification of the level values by the pitch range have been adjusted from ‘Flat, Medium, Dynamic of pitch range’ to ‘Flat, Medium (or Dynamic) of pitch range’.

Table 6: Level Values of Pitch Range of Voice Analysis Result

UX	No.	Mean	Scheffe				F	P
			Group	No.	MD	P		
F	4_1	3.140	a	4_2	-1.411	.000	67.73	.000
				4_3	-1.147	.000		
	4_2	4.551	b	4_1	1.410	.000		
				4_3	.263	.117		
	4_3	4.288	b	4_1	1.147	.000		
				4_2	-.263	.117		
I	4_1	3.132	a	4_2	-1.424	.000	69.48	.000
				4_3	-1.147	.000		
	4_2	4.556	b	4_1	1.424	.000		
				4_3	.277	.098		
	4_3	3.989	b	4_1	1.147	.000		
				4_2	-.277	.098		
E	4_1	2.625	a	4_2	-1.714	.000	98.12	.000
				4_3	-1.522	.000		
	4_2	4.339	b	4_1	1.714	.000		
				4_3	.192	.359		
	4_3	4.147	b	4_1	1.152	.000		
				4_2	-.192	.359		

* F: Functional / I : Interaction / E : Emotional

3.5. User Experience Evaluation on Rate of Voice

The average values of the voices of 1-1. Slow of rate, 1-2 Medium of rate, and 1-3. Fast of rate were compared. There was a statistically significant difference(p.value: 0.001<) in the pitch range against the A)functional/service experience, B)interactional experience and the C)emotional experience. According to the Scheff verification, as the level values of ‘slow’(group b), ‘medium’(group c), and ‘fast’(group a) were confirmed to be at all different levels, the existing level value classification was determined to be kept.

Table 7: Level Values of Rate of Voice Analysis Result

UX	No.	Mean	Scheffe				F	P
			Group	No.	MD	P		
F	5_1	3.951	b	5_2	-.692	.000	56.60	.000
				5_3	.663	.000		
	5_2	4.643	c	5_1	.692	.000		
				5_3	1.355	.000		
	5_3	3.288	a	5_1	-.663	.000		
				5_2	-1.355	.000		
I	5_1	3.920	b	5_2	-.712	.000	56.21	.000
				5_3	.629	.000		
	5_2	4.632	c	5_1	.712	.000		
				5_3	1.342	.000		
	5_3	3.290	a	5_1	-.629	.000		
				5_2	-1.342	.000		
E	5_1	3.527	b	5_2	-.790	.000	61.00	.000
				5_3	.578	.000		
	5_2	4.317	c	5_1	.790	.000		
				5_3	1.368	.000		
	5_3	2.949	a	5_1	-.578	.000		
				5_2	-1.369	.000		

* F: Functional / I : Interaction / E : Emotional

4. Conclusion

In this research, based on the researches on the literature of the classification of the auditory interface design factors, the attributes and their level values of the robot’s auditory interface design factors that can be utilized in the TTS system and SSML have been organized. A survey was conducted to investigate whether this classification would cause a significant difference in terms of the user experience. With the results from the survey, the attributes and their subordinate level values of the auditory interface design factors that affect the user experience have been suggested.

As the attributes of the voice interface design factors to measure the overall user experience with the robot, ‘Prototype of the sound effect’, ‘Gender’, ‘Pitch’, ‘Pitch range’, and ‘Rate’ have been confirmed to be useable. From the results of the subordinate level value research, only some level values have been combined. First, in the voice types classified by gender, there was no significant difference between the ‘female’ and ‘male’ voice. However, there was a difference in the evaluation for the neutral and mechanical voice that is hard to tell whether it is of female or male. Accordingly, the level values of ‘female voice, male voice, and robot-like voice’ have been adjusted to ‘human voice’ and ‘robot-like voice’. Next, in the voice types classified by the pitch range, the level values of the ‘medium’ and ‘dynamic’ pitch range voice did not have a significant difference whereas there was a significant difference in the level value of the ‘flat’ pitch range voice. Especially, the ‘flat’ pitch range voice had a negative evaluation in all types of the experiences; it seems to be useful only for certain emotional types. Thus, the three level values of ‘flat, medium, and dynamic’ pitch range have been adjusted to ‘flat, medium (or dynamic)’ pitch range level values. Consequently, two out of the surveyed subordinate level values out of 15 had been removed and after all, a total of 13 level values have been organized.

Since the auditory interface can adapt altered design factors depending on the contents of the information or emotions, the level values that are underrated now still have a high possibility of being utilized later on. In this research, only the basic information such as the platform and general phrases of feedback has been studied. In later studies, the contents of expressing the characteristics or emotions of a particular robot should be discussed. It is meaningful that this research has been conducted as a base study to systemize the factors that can be used in the robot voice interface design in terms of the user experience. As the robot appearance and the visual interface design factors have been previously studied, the behavioral interface design factors will be researched next in order to build a library that can be utilized for the user experience design based service robot design along with the results from this research.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.2017R1A6A3A11031044).

References

[1] Yoshimura T, Tokuda K, Masuko T, Kobayashi T & Kitamura T (1999), Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis, *Proceeding of EUROSPEECH* 5, 2347-2350.

[2] Lee HJ (2015). Korean Sentence Symbol Preprocess System for the Improvement of Speech Synthesis Quality. *Journal of the Korea Society of Computer and Information* 20(2), 149-156.

[3] Kim Y, Ahn S & Lee T (2017) UX Guidelines for Designing Audio Guidance of Multimedia Contents for Low-Vision - With a Focus on Sound Level Difference and Amount of Auditory Information, *Archives of Design Research* 30(1), 131-142.

[4] Asakawa C, Takagi H, Ino S & Ifukube T (2003). Maximum listening speeds for the blind. *Proceedings of the 9th International Conference on Auditory Display (ICAD)*.

[5] Kwak S, Goh T, Park K & Ahn J (2012). The Usability of a Robot and Human Empathy by Anthropomorphic Sound Feedback. *Journal of Korean Society of Design Science* 23(3), 20-27.

[6] Yang B (1997). A Study on the Human Auditory Scaling. *Voice science* 2, 125-134.

[7] Park HW, Jee SH, & Bae MJ (2016). Study on the Confidence-Parameter Estimation through Speech Signal. *Asia-pacific Journal of Multimedia Services Convergent witth Art, Humanities, and Sociology* 6(7), 101-108.

[8] Mehrabian A (1971). *Silent Messages*. Belmont, Ca: Wadsworth Publishing Company.

[9] Arons B & Mynatt E (1994). The future of speech and audio in the interface: a CHI'94 workshop. *Proceedings of ACM SIGCHI Bulletin, CHI'94 Reports* 26(4), 44-48.

[10] Garzonis S, Jones S, Jay T, & O'Neill E (2009). Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability and preference. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. *ACM*. 1513-1522

[11] Stephen AB, Peter CW, & Alistair DNE (1993). An evaluation of earcons for use in auditory human-computer interfaces. *Proceedings of the INTER-ACT'93 and CHI'93 conference on Human factors in computing systems. ACM*. 222-227.

[12] Vilimek R & Hempel T (2005). Effects of speech and non-speech sounds on short-term memory and possible implications for in-vehicle use. *Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display. International Community for Auditory Display*. 5. 344-350

[13] Chae HS, Hong JY, Jeon MH, & Han KH (2007), A Study on Voice User Interface for Domestic Appliance. *Korean Journal of the Science of Emotion and Sensibility* 10(1), 55-68.

[14] Jee ES, Jeong YJ, Kim CH, & Kobayashi H (2010). Sound design for emotion and intention expression of socially interactive robots. *Intelligent Service Robotics* 3(3), 199-206.

[15] Choi JH, Cho DU, & Jeong YM (2016). Identification of voice for listeners who feel favor using voice analysis. *The Journal of Korean Institute of Communications and Information Sciences* 41(1), 122-131.

[16] Schmitz M, Krüger A, & Schmidt S (2007). Modelling personality in voices of talking products through prosodic parameters. *Proceedings of the 12th international conference on Intelligent user interfaces. ACM.* 313-316.

[17] Cho DU (2015), Voice feature analysis of next president candidate in south korea. *Proceedings of KICS Summer Conference* 509-510.

[18] Laukka P, Juslin PN, & Bresin R (2005), A dimensional approach to vocal expression of emotion. *Cognition and Emotion* 19(5), 633–653

[19] Cho DU (2015), Voice feature analysis of CEO of the 3 major companies in south korea by applying IT Technologies, *Proceedings of KICS Summer Conference*

[20] Apple W, Streeter LA, & Krauss RM (1979), Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology* 37, 715–727.

[21] Choi Y, Jung Y, Kim Y, Suh Y, & Kim H (2018), An end-to-end synthesis method for Korean text-to-speech systems, *Phonetics and Speech Sciences* 10(1), 39-48.

[22] Lee JS (2003), Design and Implementation of a Speech Synthesis Markup Language Processing System for Korean, *Dissertation ,Graduate School of Sungshin Women's University*.

[23] W3C (2010), Speech Synthesis Markup Language (SSML) Version 1.1, available online: https://www.w3.org/TR/2010/REC-speech-synthesis11-20100907/ last visit: 01.02. 2018

[24] Chung SE & Ryoo HY (2018), Aural and Behavioral Factor Research for Robot Design, *Proceedings of International Journal of Multimedia and Ubiquitous Engineering* 13(2), 1-6.