Spatial voice menu and head gesture interaction system for a wearable computer

Taiki Saito*
Tokyo Metroplitan University

Yasushi Ikei†
Tokyo Metropolitan University

Koichi Hirota‡
The University of Tokyo

Michitaka Hirose§
The University of Tokyo

Abstract

In this paper, we describe a voice menu/head gesture interface, the vCocktail+, built for efficient interaction with a wearable computer. This method is devised to reduce the length of the serial presentation of voice menus by introducing spatiotemporal multiplexed voices with enhanced separation cues. The voice menu items were presented fast only within 2 seconds, where the system allows the user to achieve a correct hearing ratio of 99.7 % in the case of four-item multiplexing. The head motion (gesture) input system was installed to select the menu presented by a spatial sound. Six kinds of head gesture were recognized by processing the data from 6-dof acceleration sensor attached to the headphones. The recognition performance was sufficiently reliable to be used for a fast input system. The vCocktail+ system was applied to control a digital multimedia player iTunes to demonstrate the utility of the system. The result of a preliminary experiment suggested that the vCocktail+ allowed the user to interact with a wearable computer fast enough just as a GUI interaction with a large visual screen and a mouse on a non-wearable computer.

Index Terms: H.5.2 [User Interfaces]: Input devices and strategies—Interaction styles;

1 Introduction

Wearable is a new style of computing, and it is thought to give us efficient and useful experience anytime anywhere in our daily life. One of the characteristics of wearable computing is that it provides quick interaction to relevant information to support and augment the main task of the user at the moment. So, the displayed information needs to be grasped by the user quickly while control load of the wearable computer itself should be minimized. Current human interface of wearable computers depends largely on visual interaction and manipulation of an input device by hand, which often interferes with the execution of the main task.

The final goal of this research is to establish an efficient voice and gestural interaction method for wearable computers. Here, in this paper we introduce an efficient voice menu presentation and head-gesture input system called vCocktail+ for a quick interaction with a wearable computer. The conventional voice menu presentation is a sequential readout of menu items, which takes a long time and gives an undesirable redundant impression to the users. The vCocktail+ system enables efficient information transfer with reduced presentation time by introducing a temporally multiplexed voice menu presentation. For the separation of menu items, a spatial sound localization technique is employed to assure stable listening.

*e-mail: saitou@krmgiks5.sd.tmu.ac.jp
†e-mail: ikei@computer.org
‡e-mail: hirota@media.k.u-tokyo.ac.jp
§e-mail: hirose@cyber.t.u-tokyo.ac.jp

20th International Conference on Artificial Reality and Telexistence (ICAT2010), 1-3 December 2010, Adelaide, Australia
ISBN: 978-4-904490-03-7 C3450

2 Related work

A simple technique to shorten the time to present information by voice is to increase the rate of speech. Both inhibiting pitch change of speech (frequency change) and still augmenting the speed of speech are possible by digital processing, however the ease of hearing is markedly degraded if the speed of speech is increased. Another technique [1] could shorten the long speech by removing (shortening) the space between words, however it is not applicable to read out many menu items for the purpose of interaction with a computer.

For improving the browsability of auditory information, support in the simultaneous listening of multiple voices and the transfer of attention between them is required, so that introducing the spatial property of voice is expected to be effective. A binaural system, stereo recording through a dummy head to produce a spatial sound with headphones, has been used for reproduction of sound in a real space [2]. However, it was difficult to add, during replay, the different spatial properties to the recording. Recently, localization of the sound source at an arbitrary position in real time has been easily performed due to improved processor performance [3]. Spatialized audio has been installed in Virtual Reality (VR) displays [4].

AudioStreamer [5], for example, introduced a method that enabled efficient listening of multiple speeches by localizing three sound sources in a space simultaneously. Dynamic Soundscape [6] rotates up to four voices in a space to enable multiple portions of a long speech file to be heard simultaneously. It also localizes the topics in particular directions to map the timeline to a space for providing quick access. A 3-dimensional auditory radial pie menu (in non-speech audio) around the user's head with head-gesture selection was also investigated [7] as an interaction technique for wearable devices.

Presenting and selecting plural menu items by a pull-down (pop-up) menu used in GUI is a very fast and effective interaction method; however, an efficient equivalent system in voice interaction has not been established yet. Conventional voice menu presentation requires a long time to read out the menu items. The user's attention resource is occupied inefficiently with it, and the cognitive load during the interaction is large. This greatly decreases the advantage of the hands-free and eyes-free features of the voice interface.

In a voice menu presentation research focusing on the GUI menu [8], a system that uses the stereo volume difference method in which the accuracy of hearing is compared with 3-times-faster playback has been proposed. Three-voice simultaneous presentation achieved an 88 % correct answer ratio that was not sufficient for the interaction. It seemed that sound images were not distinguished effectively only using the level difference between two speakers since it did not localize the sound image out of the user's head. In addition, complete simultaneity of voices is hard to perceive for humans.

A hand and head gesture was used for the input to the voice menu in the wearable computing system [7] in which the usability may be increased and the interference to the main task should be alleviated. However, many issues remains for the user load and the operation time. In addition, 20% false detection of head gesture was not sufficiently small to work for usual input system.

An input system based on head gestures for impaired people was investigated [9] as a part of an electric wheel chair control interface. The system was built for a severely disabled people who was not able to move by the hand a control stick of a electric wheel chair. Head gestures were adopted alternatively. A stereo-camera mounted on the chair detected the head gesture to use it for a command input to the controller. Recognition of head gestures by computer vision are investigated by others [10],[11] based on a recent fast cpu. However, a highly accurate gesture recognition is not necessarily feasible in the wearable use since the user's posture is diverse as well as the camera position, lighting, background objects, walking motion blurs are inevitable noise sources.

In this paper, we introduce a multiplexed presentation and hands/eyes-free input method, vCocktail+, that presents up to four voice menu items in a space around the head and allows the user to select (input) the item to interact with a wearable computer.

The rest of this paper is organized as follows. Section 3 describes overview of the system design. First, the technique to present fast voice menus is introduced to show how the parameters of voice presentation was obtained. Then, the basic idea of selecting menus by head gestures is presented. Section 4 discusses the design of head gesture recognition algorighm and its evaluation regarding the pitch motion detection. In Section 5. the application example of the vCocktail+ system to the multimedia player software, the iTunes. The preliminary evaluation based on a model task is presented. The last section summarizes the results and shows future works of this study.

3 Sound localization/head gesture input system

The vCocktail+ system is a hands-free/eyes-free interaction system for a wearable computer. It localizes the voice menus in a horizontal plane around the user's head[1] by using headphones. The voice sounds are multiplexed in time for fast and efficient menu presentation. The presented menu items are pointed (and activated) by head motion (gesture) to perform the menu's function. In addition, two gestures of nodding and upturning head motions are also available for special input. The system provides a new method to interact with a wearable computer.

3.1 Voice menu display system of vCocktail+

The vCocktail+ introduced two multiplications as shown in Fig. 1 for quick voice menu presentation. The voice menu items are sounded overlapping the words with a short onset delay time to shorten the total time to playback the menus. In addition, the menu-item voices are played in four separate directions to alleviate independent hearing of each item. In addition, the voices are progressively attenuated (the playback volume of a menu voice is attenuated to zero at the end of the word) so that the hearing separation between words is easily done[12]. The parameters of the presentation were determined by the result of evaluation experiment to cover 99.7 % correct hearing as follows. The number of menu items was four; the number of directions to place the voice images were four; the onset delay of sounding of a menu word was 0.3 [s].

The sound image localization system is based on a single common (not personalized) HRTF with headphones. The localization accuracy was measured first to find the optimal design to utilize spatial characteristics. Absolute direction perception error was 20 degrees on average ignoring front/back mistakes. Although this value needs to be improved by acquiring an HRTF that better fits the user, we adopted a 60-degree interval only in the frontal direction for multiple-sound localization. This interval angle was expected to be large enough for the user to determine the sound direction since the relative directional judgment is easier than the absolute judgment.

[1]The system actually performs three-dimensional localization. However, in general, the perception of elevation of a sound source is not accurate with a non-personalized HRTF.

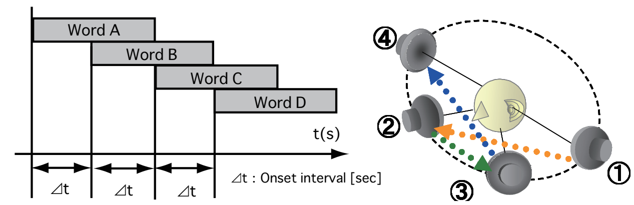

Figure 1: Temporal and spatial multiplexed presentation of voice menu items.

As design parameters, the number of voice menu words, onset interval time between the words, attenuation of voice, spatial sequence alternatives, and whole-list/target-word listening conditions were provided in the experiments to investigate the characteristics. The result showed that a large onset interval, a small number of menu items, using attenuating voices, and cross-type localization are effective for error ratio reduction in discriminating a multiplexed voice menu.

A small onset interval and a large number of menu items are preferable for a quick presentation of voice menus. However, if we prioritize the lower perception error ratio, the following design criteria are obtained from the data that demonstrated the relation between error ratios over the alternative conditions. If we need a correct answer ratio higher than 99.7 %, a 0.3s onset interval for up to 3 menu items, and a 0.4s onset interval for a 4-item menu are necessary without spatial localization and no attenuating voices. If we introduce the cross-type spatial localization procedure (as shown in Fig. 1), a 0.3s onset interval meets the correct answer ratio for the 4-item menu. Moreover, the attenuating voice enables a 0.2s onset interval even for the 4-item menu. These values indicate basic characteristics of the voice menu multiplexed presentation method vCocktail+ proposed in the present paper.

3.2 Menu selection by head gestures

The user selects a menu item by rotating the head toward the orientation where the target item is heard. This rotation gesture allows the user to point a target item intuitively from four items presented around the head. The menu items on the left and right sides are selected by a roll motion that tilts the head sideways, while the items in front of the face are selected by a yaw motion to place the item right in front of the user's face. In addition, not related to the spatially presented menus, a pitch motion is also detected.

The unit motion of the gesture is to rotate head to a particular direction from the neutral position (usually to face forward) and then rotate back to the initial position quickly. The speed of rotation is relatively high during the motion to move and return the head to the neutral position in a short time, and the amplitude of rotation is small although they were not designated. The detection of gestures of roll, yaw and pitch motions is performed by a pattern processing algorithm.

4 Gesture motion recognition and evaluation

The head gestures used in this system are six kinds, the roll motions to right and left, the yaw motions also to right and left, and the pitch motions to up and down. The pitch and yaw motions are often used in everyday life such that nodding indicates agreement or decision while shaking head means denial. These are intuitive motions for us in social circumstances.

We discuss the design of pitch motion here since other motions were already implemented in the same way.

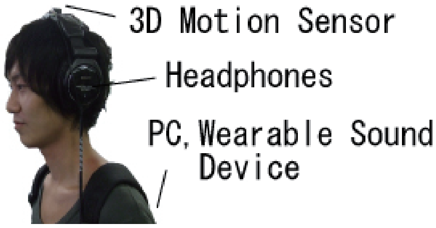

Figure 2: System devices of vCocktail+

4.1 Gesture detection system

The rotation motion of the head was measured by 3-axis angle sensor (MDP-A3U9, NEC Tokin). The sensor was placed at the top of a head belt of the headphones that were used in the vCocktail+ system as shown in Fig. 2. The sensor itself is small enough to be installed in a small earphones in the future implementation.

The three axises discussed in the present paper are those fixed to the head coordinates as follows.

Roll axis: an axis normal to the face.
Pitch axis: an axis that goes through both ears.
Yaw axis: an axis that goes from the neck to the top of head.

The motions around these axises are written as a roll motion, a pitch motion, and a yaw motion hereafter. The range of the sensor is ±180 [deg] for the roll and yaw axises, ±90 [deg] for the pitch axis. The sampling rate was 125 [Hz].

The rotation signal observed during a straight walk was about ±20, ±30, ±35 [deg] for yaw, pitch and roll motion, respectively, from the result of a preliminary experiment.

4.2 Pitch gesture recognition

The ratio of angular velocities of outward and return was used to define a single gesture of input to avoid false recognition by a noise from the body motion. In addition, the ratio of the amounts of angular-velocity changes (integrals of angular acceleration) of outward and return motions, A, described by Eq. (1) was incorporated as a criterion. Also the time in gesture before return, B in Eq. (2) was introduced in the judgment.

$$A = \frac{\overline{\delta\omega_2}}{\overline{\delta\omega_1}} < \alpha \qquad (1)$$

$$B = m - l < \beta \qquad (2)$$

where $\overline{\delta\omega_1}, \overline{\delta\omega_2}$ are obtained from Eq. 3 and 4.

$$\overline{\delta\omega_1} = \begin{cases} \frac{\sum_{i=l-24}^{25} |\omega_i - \omega_{i-1}|}{25} & (l-k \geq 25) \\ \frac{\sum_{i=k}^{l-k} |\omega_i - \omega_{i-1}|}{l-k} & (l-k < 25) \end{cases} \qquad (3)$$

$$\overline{\delta\omega_2} = \begin{cases} \frac{\sum_{j=m}^{25} |\omega_j - \omega_{j-1}|}{25} & (n-m \geq 25) \\ \frac{\sum_{j=m}^{n-m} |\omega_j - \omega_{j-1}|}{n-m} & (n-m < 25) \end{cases} \qquad (4)$$

where variable k, l, m, n are the time the angular velocity crosses the values of ±40 [deg/s] as shown in Fig. 3. The unit of the variables is 8 [ms]. $\overline{\delta\omega_1}, \overline{\delta\omega_2}$ are the average absolute angular acceleration within 200 [ms] before l and after m, respectively. The ranges over 40 [deg/s] were included in the algorithm considering the magnitude of a noise from walking in addition to the velocity range the user selected.

To determine the threshold α of the ratio A, the false rejection rate (FRR) and the false acceptance rate (FAR) based on the data from preliminary experiment were introduced.

For the upturn gesture, the value of α was determined 1.02 where FRR=5.3%, FAR=5.8%. For the nodding gesture, the α=1.12 where FRR=8.4% and FAR=22.8%. In addition, the threshold β of return time of a gesture was determined as β=50 and 55 (400 and 440 [ms]) for the upturn and nodding gestures where the FRR=0.0% for the both.

The criteria of pitch motion recognition are summarized as follows:

(a) The angular velocity of head motion exceeds the threshold. The threshold values were 60 and 40 [deg/s] for the outward and return motions, respectively.

(b) After the angular velocity exceeded 40 [deg/s], the ratio 'A' is below 1.02 and 1.12 for upturn and nodding, respectively.

(c) The time B was less than 400 and 440 [ms] for the upturn and nodding, respectively.

4.3 Evaluation experiment

A false recognition and a detection loss during the pitch gesture recognition were measured for the evaluation of the design. The procedure of the experiment was as follows.

The participant wears headphones with a sensor and hears the voice instruction read by synthesized voice that orders to do either upturn or nodding. Then the participant performs designated head gesture. This unit, a single trial, was repeated 50 times for nodding and upturn gestures both in sitting and walking conditions. In a walking condition, the participants walked straight at an ordinary speed for themselves. The participants were six males and one female students (average age of 22.3) with normal auditory sense and physical capabilities.

4.4 Results and discussion

4.4.1 False recognition

The amount of false recognition occurred in the experiment was shown in Fig. 4. It was 0 % and 1.43 % for upturn and nodding gestures, respectively, in a siting posture. Those in straight walking were 0 % and 2.57 %, respectively. The rate of false recognition was higher in nodding than upturn since the threshold α of criterion (b) of pitch motion detection was determined with higher FAR of 22.8 % in nodding than upturn with 5.8 %. It looks that criteria (b) and (c) were successful since false recognition was 0.0 % for upturn gesture in both sitting and straight walking postures.

4.4.2 Detection loss

In a sitting posture, the detection loss was 2.0 % and 11.4 % for upturn and nodding gestures, respectively. In straight walking, it

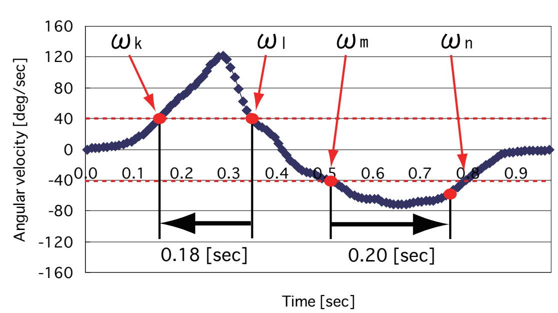

Figure 3: Angular velocity profile of a gesture

was 3.7 % and 8.9 % (Fig. 5) .

The detection loss was more observed in nodding than upturn gesture in the both postures of sitting and walking. This is because that gesture recognition was less sensitive for nodding than upturn since the nodding FRR that determined α in the criterion (b) was as high as 8.4 % in contrast to 5.3 % of upturn.

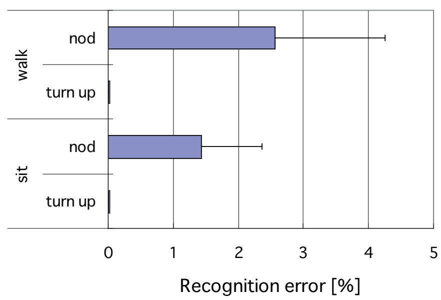

Figure 4: False recognition (Error bar indicates SEM)

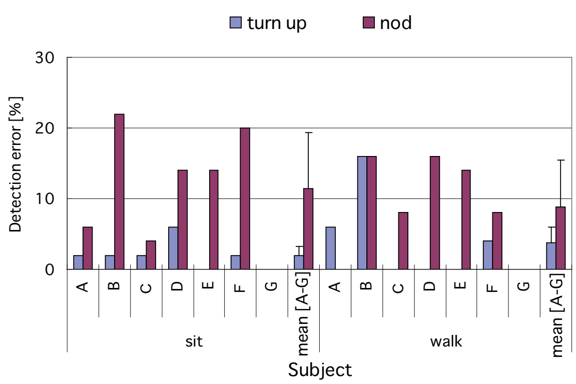

Figure 5: Detection loss (Error bar indicates SEM)

5 Evaluation of implementation of vCocktail+

vCocktail+ provides a rapid interaction style for a wearable computer even during the task that requires hands. In general, the purpose of wearable computer operation is either related or indifferent to the main task that is going on. The main task is distinguished from others in that it occupies user's attention longer than others.

In this chapter we demonstrate that operation of a wearable computer with vCocktail+ does not much affect the main task performed concurrently. More specifically, the vCocktail+ is applied to the control of a digital music player that is generally used by many people doing other tasks. A task completion time was investigated while the control of the music player interrupted the task sporadically.

5.1 Implementation to iTunes

The vCocktail+ system that included six-channel input was applied to Apple Inc.'s multimedia player, iTunes run on a desktop computer. The menus were organized in a multi-level structure, and the items in a level were presented with a multiplexed voices. The

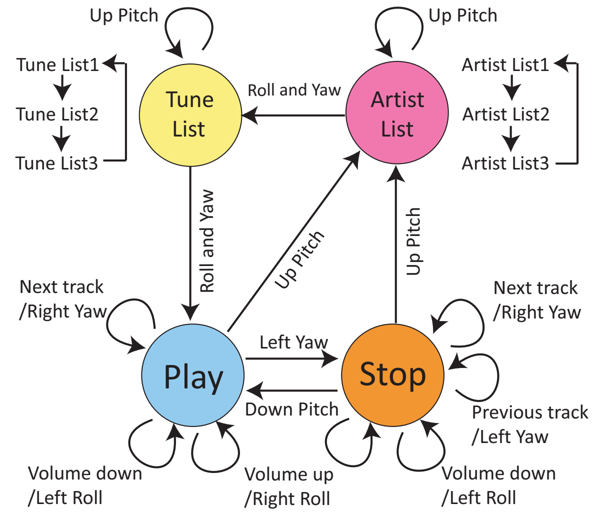

Figure 6: State transition diagram of iTunes control interface

voice menus were presented in a form with four words and 0.3 [s] onset asynchrony, attenuating amplitude, and the cross type spatial arrangement. The four menu items in a particular level were presented to the user to control the iTunes as designated, then the head motion input was performed. The menu presentation at the level was repeated if the menu item was not selected after a 0.7 [s] delay. The menu presentation of a single level (four items) ends within about two seconds. Twelve artists and four songs for each artist were installed in the iTunes. (Forty-eight songs in total were used.)

The vCocktail+ to interface iTunes has controls for artist selection, song selection, volume adjustment, scan within the list in a particular level, and start/stop of playback. Figure 6 shows a state transition diagram of the interface program. An upturn gesture at the *Stop* state presents the names of four artists from an artist list. Then the user selects one of the four by a head-gesture, otherwise if the user turns up again, the next page of the artist list is presented. If an artist is selected, then a song list is presented by a multiplexed voice menu. If the user selects one of the songs, then the song is played.

At the *Play* state, yawing to the right skips to the next song while yawing to the left stops playing. At the *Stop* state, yawing to right and left moves focus on the next and the previous song, and then a nodding plays the song. Either at *Play* or *Stop* state, rolling to right and left increases and decreases the volume, respectively.

5.2 Preliminary evaluation

The utility of vCocktail+ was evaluated with the following procedure. The participant changes the song as designated by the experimenter, concurrently performing a drill of calculation as a main model task. The main model task is to fill the cells of a table with answer figures adding two numbers on the top row and the left column as shown in Fig. 7. The task was provided on a PC screen, and the participant inputs answers by a keyboard.

This is a model of swift computer operation (to change a song by a head gesture) during the task that requires mental concentration. A main task usually takes a large part of user's attention. This model task is intended to give the participant a high cognitive load with visual attention. The experimenter gave an indication to the participant to change a song five times every one minute. The indication was given to the participant each time by showing a paper with the name of a song on it.

The reference condition is that the participant changes by a mouse a song on iTunes displayed on the same computer screen that requires to change the window from the Excel to the iTunes. In addition, a control condition in comparison was prepared that the participant performed the drill calculation without the operation of the iTunes. The number of participant was six.

Figure 8 shows the average time for each condition. The time to finish the task of filling cells was 9.0 [min] and 9.4 [min] for the vCocktail+ operation and the GUI operation, respectively. All the participant could play the songs designated by the experimenter without fail. This result of completion time indicates that the vCocktail+ operation was as efficient as the GUI operation. The difference between the two was not significant, although *no task* condition was significantly shorter. ($F(2,15)=7.69$, $p=5.0\times10^{-3}$)

In addition, the mean time of single song selection was measured in case that the operation involved the list change. When selecting from the first list of artist, it took 8.7 [s] while from the second and third lists it took 11.5 [s] and 16.1 [s], respectively. The increase of the time is a total duration of an upturn gesture to proceed to the next list, listening the menu, and selection of an item that was from 2.8 to 4.6 [s]. This value is an estimate time for a single selection.

In addition, though not having sufficient amount of data, the quality of the main task looks better with the vCocktail+ than using the GUI since the result of the calculation was more accurate with vCocktail+. Figure 9 shows the error ratio of the main task (calculation error). A tendency was observed that the interference to the main task was less with vCocktail+ that require no on-screen operation than the GUI operation that produced visual disturbance when commanding the iTunes, although the difference was not significant ($F(2,15)=1.6$, $p=0.23$).

	A	B	C	D	E	F	G	H	I	J	K
1	+	50	93	9	15	69	98	84	95	81	37
2	85	135	178	94	100	174	183				
3	3	53	96	12	18	73	101				
4	18	68	111	27	33	87	116				
5	78	128	171	87	93	147	176				
6	40	90	133	49	55	109	138				
7	46	96	139	55	61	115	144				
8	31	81	124	40	46	100	129				
9	41	91	134	50	56	110					
10	48	98	141	57	63	115					
11	89	139	182	98	104	158					

Figure 7: Calculation drill of 100 cells

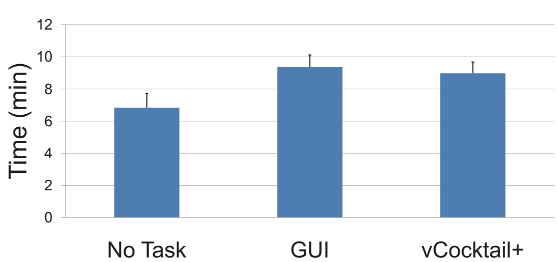

Figure 8: Completion time (Error bar indicates SEM)

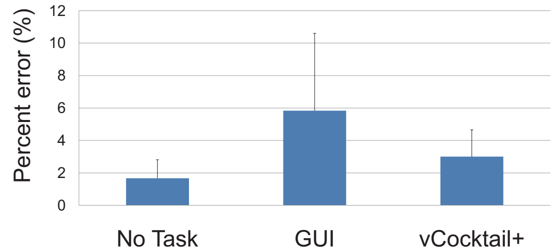

Figure 9: Calculation performance (Error bar indicates SEM)

6 Discussion

In the present study, a multiplexed voice menu presentation and head gesture input system, vCocktail+ was introduced to efficiently utilize the voice menu and hands free selection to interact with wearable computers. This design reduces the length of speech presentation and allows quick menu selection without hand. On the basis of user study, the characteristics of vCocktail+ were investigated and the parameters of the design were obtained.

The sound image localization was based on a HRTF that was essentially dependent on individual conditions related to hearing. Thus, current implementation would not be optimal for each participant, however even with that condition, the vCocktail+ system exhibited sufficient separation in hearing. To measure individual HRTF requires practically certain amount of cost which may decreases the value of relative usefulness of the system.

In this research we used a certain set of parameters that is easily perceivable for many participants. However, it is still very fast compared to the sequential playback of the menu items (names of artists or songs in this case). The participant of the experiment easily captured the designated menu item and could select it by head gestures successfully. The completion time added by control of the iTunes was almost the same between vCocktail+ and GUI operations. As the participants had sufficient skills for operating GUI menus, they performed very fast change of songs with GUI. The vCocktail+ operation time was even less than GUI. The result clearly demonstrated the practical utility of the system.

7 Conclusion

The vCocktail system was first developed to present voice menus to the user on headphones efficiently in a shorter time than a usual sequential playback of menu items. The system locates voice menu images in a 3D space with enhanced separation cues of an onset delay, a cross-type spatial arrangement, and an attenuating sound. The design worked successfully for four menu items. Then, an input system was built to select the menu items to produce a command to a wearable computer. This new method utilizes a relatively small rotation of the user's head to the direction of the spatial menus that enables a hands-free and eyes-free menu selection input system vCocktail+.

In the present paper, the design of a head gesture recognition system was described, then an application of the system to the music player, iTunes was evaluated as a practical implementation example. The characteristics of head gesture motions were measured to obtain parameters of the gesture recognition system. Then the values were determined for a pitch motion as well as yaw and roll motions. False recognition was 0.0% for the upturn gesture, and 1.43% for the nodding gesture. The input system with least false recognition of pitch motion was successfully implemented to the system. There were unrecognized gestures of 6.5% on average, however the authors consider that the recognition accuracy could be improved by the proficiency of the user to the system and also with the individual algorithm (parameter) optimization.

The evaluation of vCocktail+ interface implementation was investigated regarding the iTunes operation during a calculation task. The calculation completion time was almost equivalent to that using GUI with a mouse on a PC screen, which suggested that the vCocktail+ interface could minimize interference to the task as generally fast and effective GUI settings. In addition, the quality of the task looked better with vCocktail+ operation than GUI. This quality measure is still preliminary, which should be investigated further about the difference of cognitive interference between vCocktail+ and GUI.

The future work includes a multistage menu structure design that provides more than four menu items that will be installed in a practical system. Thus, the timing for repeated presentation of items and for selection of items needs an appropriate design. In addition, we plan to install the vCocktail+ system into other practical application such as an email client or a scheduler. That will require investigation of consistency of a menu hierarchy among particular applications. We continue to investigate the utility of vCocktail+ system.

Moreover, the tuning of the system element still needs improvement. The head-gesture recognition algorithm should be optimized to individuals, and the adaptive control of recognition parameters might be needed to fit to the posture (motion state) of the user. These would contribute to the accuracy of gesture recognition and detection ratio.

REFERENCES

[1] B. Arons. SpeechSkimmer : Interactively Skimming Recorded Speech. Procs. of UIST '93, 1993.

[2] Y. Yamasaki. Human Auditory Characteristics and Sound Field Control. JIPSJ, Vol. 36, No. 3, pp. 244-252, 1995.

[3] E. M. Wenzel, D. J. Miller, J. S. Abel. A Software-Based System for Interactive Spatial Sound Synthesis. ICAD2000, pp. 151-156, 2000.

[4] T. Ogi, T. Kayahara, M. Kato, H. Asayama, M. Hirose. Immersive Sound Field Simulation in Multi-screen Projection Displays. Immersive Projection Technology and Virtual Environments 2003 (IPT/EGVE 2003), pp. 135-142, 2003.

[5] C. Schmandt, A. Mullins. AudioStreamer: Exploiting Simultaneity for Listening, ACM CHI95, pp. 218-219, 1995.

[6] M. Kobayashi, C. Schmandt. Dynamic Soundscape: Mapping Time to Space for Audio Browsing. CHI'97, pp. 22-27, 1997.

[7] Brewster, S., Lumsden, J., Bell, M., Hall, M., and Tasker, S.：Multimodal 'Eyes-Free' Interaction Techniques for Wearable Devices；ACM CHI, 2003. Fort Lauderdale, FL: ACM Press, Addison-Wesley, pp. 463-480(2003).

[8] N. Kawamoto, et al. An Evaluation of Methods for Menu Presentation using Multiplex Voices. Technical report, IPSJ, HI59-7, 1995.

[9] YODA Ikushi, TANAKA Junichi, KIMURA Yusuke, RAYTCHEV Bisser, SAKAUE KatsuhikoI, NOUE Takenobu, Non-contact Non-constraining Head Gesture Interface System for Electric Wheelchair, IEICE transactions on information and systems, J91-D(9), 2247-2255, 2008 (in Japanese)

[10] Fumihiro Matsumoto, Youhei Egawa Yasuyuki Kono, Masatsugu Kidode, Recognition of Head Motions using IR Image Sensor, IPSJ, 2001-HI-932001(38), 15-22, 2001 (in Japanese)

[11] R.E. Kaliouby, P. Robinson：Real-time Inference of Complex Mental States from Facial Expressions and Head Gestures；Real-Time Vision for HCI, Springer-Verlag, pp. 181-200(2005).

[12] Yasushi Ikei, Hitoshi Yamazaki, Koichi Hirota, Michitaka Hirose, vCocktail: A wearable-oriented multiplexed voice menu presentation method, Trans. Human Interface Society, Vol. 7, No. 4, 2005 (in Japanese)