Multiple Information Communication in Voice-based Interaction

Muhammad Abu ul Fazal (✉) and Muhammad Shuaib Karim

Department of Computer Sciences
Quaid-i-Azam University, Islamabad, Pakistan
fazalsidhu@yahoo.com, skarim@qau.edu.pk

Abstract. Ubiquitous Computing has enabled users to perform their computer activities anytime, anyplace, anywhere while performing other routine activities. Voice-based interaction often plays a significant role to make this possible. Presently, in voice-based interaction system communicates information to the user sequentially whereas users are capable of noticing, listening and comprehending multiple voices simultaneously. Therefore, providing information sequentially to the users may not be an ideal approach. There is a need to develop a design strategy in which information could be communicated to the users through multiple channels. In this paper, a design possibility has been investigated that how information could be communicated simultaneously in voice-based interaction so that users could fulfil their growing information needs and ultimately complete multiple tasks at hand efficiently.

Keywords: Voice-based Interaction, Multiple Information Broadcast, Multiple Voices, Information Design

1 Introduction

In this information age which is highly influenced by technology, people have many computing devices and associated interaction modes to fulfill information needs and perform desired tasks conveniently from anywhere [1]. For example, mobile telephony has become an essential tool [2] that humans carry with them almost all the time. It is playing a significant role to access information on the go by either interacting visually or by using voice-based interaction. A voice-based interaction is a mode where users are provided with the facility to interact with the system using 'voice'.

The motivation of using voice to interact with the system is an old concept which can be associated with Ali Baba's 'Open Sesame' and earlier science fiction movies. The voice-based interaction method enables the user to interact with the system in immersive environment [3]. *Voice User Interfaces (VUI)s are user interfaces using speech input through a speech recognizer and speech output through speech synthesis or pre-recorded audio* [4].

The voice-based interaction enables users to conveniently interact with the system in the hand busy or the eye busy environment. This mode is also an alternative for the visually impaired users to interact with systems. According to world health organization [5], it is estimated that there are 285 million people who have visual impairments.

Humans are capable of listening and comprehending multiple information simultaneously through their auditory perception, but presently, voice-based interaction design is providing sequential interaction approach which is somehow under-utilizing the natural human perception capabilities [6]. Since the voice-based interaction is sequential therefore system provides only a limited amount of information each time, which makes it hard for the users to get an overview of the information, particularly in the case of assistive technology used by the visually impaired users [8].

One of the main goals in information design is rapid dissemination with clarity. Since users have growing information needs[9] , therefore, it must be efficiently designed, produced and distributed, so that users could quickly interpret and understand it using their auditory capabilities. If we critically look contemporary implementations, then there arises a question whether present sequential information designs are utilizing the human auditory capabilities in voice-based interaction effectively & optimally or not?

Rest of the paper is organized as follows. Next section describes Literature Review. The limited exploitation of human auditory perception is discussed under the section Auditory Perception's Exploitation Gap. The concept of communicating multiple information using multiple voice streams is discussed under Motivating Scenario section. Then, based on the motivating scenario, an experiment is described in detail under Experiment Section. Conclusion and future work is discussed under Conclusion and Future Work section.

2 Literature Review

Voice-based interaction s often used in todays' computing era that is ubiquitous in nature. Over the Web, efforts have also been made to realize voice-based user agents such as voice-based Web browsers under the Spoken Web Initiative [10]. That would benefit people who are unable to conveniently use the internet due to various reasons including low literacy, poverty, and disability.

There are many other uses of voice in system interaction like in e-learning system, the aural access is being provided as a complimentary method to the visual-only content [11]. Numerous interactive voice response applications are developed to provide important information to the targeted users, particularly the illiterate users. Interactive voice application 'Avaaj Otalo' [12] provides essential information to the low literate rural formers. Using this application, farmers can ask questions, and browse stored responses on a range of agricultural topics.

From the user side, Lewis suggested that user system interaction performance is affected by the users' characteristics like physical, mental, and sensory abilities [13]. For voice, the main sensory capability is auditory acuity. The American Speech-Language-Hearing Association has identified central auditory process as the auditory system mechanisms and processes responsible for the following behaviors [14]:

- Sound localization and lateralization, i.e. users are capable of knowing the space where sound has occurred
- Auditory discrimination, i.e. user has the ability to distinct one sound from another

- Auditory pattern recognition, i.e. user is capable of judging differences and similarities in patterns of sounds
- Temporal aspects, i.e. user has abilities to sequence sounds, integrate a sequence of sounds into meaningful combinations, and perceive sounds as separate when they quickly follow one another
- Auditory performance decrements, i.e. user is capable of perceiving speech or other sounds in the presence of another signal
- Auditory performance with degraded acoustic signals, i.e. user has the ability to perceive a signal in which some of the information is missing

Humans are able to listen to the sound whose frequency varies between 16 Hz to 20KHz. In order to perceive the two frequencies separately the width of the filters, also called 'critical band', determines the minimum frequency spacing. It would be difficult to separate two sounds if it falls within the same critical band. Besides frequency, other important perceptual dimensions are pitch, loudness, timbre, temporal structure and spatial location.

Humans are capable of focusing their attention to an interested voice stream if they perceive multiple information simultaneously as reflected in experiment discussed in this paper. For attention user adopts two kinds of approaches, one is overt attention and second is covert attention. In covert attention the region of interest is in the periphery. So, if a user is listening multiple voices, he may be interested in focusing the voice provided to him in the periphery. The regions of interest could be four to five [15]. For selection and attention in competing sounds, it is an important consideration for the listener that how auditory system organizes information into perceptual 'streams' or 'objects' when multiple signals are sent to the user. In order to meet this challenge, auditory system groups acoustic elements into streams, where the elements in a stream are likely to come from the same object [Bregman, 1990].

A few research studies exist on communicating information using voice simultaneously. The experiments have been conducted particularly in the case of visually impaired persons. According to Guerreiro, multiple simultaneous sound sources can help blind users to find information of interest quicker by scanning websites with several information items [16]. Another interesting work where Hussain introduced hybrid feedback mechanism i.e. speech based and non-speech based (spearcon) feedback to the visually impaired persons while they travel towards their destination [17]. The feedback mode alters between above two modes on the basis of the frequency of using the same route by the user and representativeness of the same feedback provided to the user. The experiment conducted by the researcher reflects that hybrid feedback is more effective than the speech only feedback and non-speech only feedback. In another study for blinds to understand in a better way the relevant source's content, Guerreiro and Goncalves, established that use of two to three simultaneous voice sources provide better results [18]. The increasing number of simultaneous voices decreases the source identification and intelligibility of speech. Secondly, the author found that the location of sound source is the best mechanism to identify content.

Above mentioned behavioral characteristics and research work suggest that human auditory perception has remarkable capabilities which are somehow not fully exploited

in the contemporary implementations of the voice-based human-computer interaction, particularly for sighted users.

3 Suggested Improvements

Contrary to the voice-based interface, the visual interface provides multiple information to the user in many ways such as using overlays [20]. Figure 1 is a Facebook wall of a user where multiple information is being communicated simultaneously. One overlay is providing the facility to view the messages being received in the conversation. Another overlay at the top is showing notifications. The right side pane is showing the activities of fellows. The left side pane is displaying his favorites and other useful stuff. And as soon as the mouse is rolled over to the text Farrukh Tariq Sidhu the preview of Farrukh's wall gets displayed in another layer. If the user is interested in the additional information provided through overlay the user may go with it otherwise ignore the overlay and would stay on the main screen.

The same design technique may be adopted in voice response system to communicate multiple information simultaneously because auditory system is capable of performing filtration of received sounds and allows the user to ignore the irrelevant noise and concentrate on important information [7].

In next section, we have discussed a scenario where multiple voice streams can help users to fulfill their information needs.

Fig. 1. An example of overlay in Graphical User Interface

4 Motivating Scenario: Listening Multiple Talk Shows

Daily, in prime time i.e. 8:00 pm to 9:00 pm various news channels air talk shows focusing different topics with different participants and hosts. People working in offices in evening or night shifts usually watch these programs live using video streams provided

by news channels, if they are free to do so at the desk. If users are busy in official work or their computer screen is occupied for another task they may prefer to listen to live audio stream from relevant channels website.

Users may be interested in listening to more than one talk shows at the same time. For an example, a person is interested in listening to the talk show 'Capital Talk' at 'Geo News' and also interested in listening to 'Off the Record' played on another channel 'ARY News'. The first talk show Capital Talk is discussing the current situation arisen due to the heavy floods whereas the second program is discussing the political scenario in Pakistan. The user is mainly interested in listening to the program discussing the political situation but also wants to know the key facts or get an overview about the flood situation being discussed in Capital Talk show.

In this perspective, user's multiple information needs may be fulfilled using multiple information communication simultaneously. In this case, information seeking could be possible in a way that a user opens two web browsers and play both the audio streams simultaneously and listens both the program in parallel. This could be challenging and-complex task for the user. The listening complexity may be reduced by keeping one streams volume low but audible and keep the main programs voice normal so that user could keep the focus on primary program. The high volume is expected to help him to keep the focus on the main program while the secondary low volume would continuously give him the feedback or glimpse that what is going on in the other program. Using this approach user might not miss the content of the program in which user is mainly interested and also get an overview of the secondary program.

This approach of playing multiple audio streams in parallel may be extended to more than two audio streams where information like a commentary on cricket match could also have listened.

In order to meet this challenge, we have framed following three research questions which we are trying to answer by conducting a series of experiments.

- How many voice streams can optimally be played to users for communicating information simultaneously?
- What could be the optimal auditory perceptual dimensions' settings of streams for better discrimination between voice streams?
- What scenarios / challenges users can face in multiple information communication?

5 Experiment

In this experiment, an audio bulletin was built wherein the voice-based information was designed in a way that two different voice streams (using female and male voices) were played simultaneously. The female voice stream was of BBC Urdu's renowned TV presenter 'Aaliya Nazki' and reporter 'Nasreen Askri' whereas male voice was of another BBC Urdu's TV presenter 'Shafi Taqqi Jami'.

5.1 Experiment Design and Settings

In order to build an audio bulletin, two different video bulletin of BBC Urdu's program 'Sairbeen' were selected. Sairbeen is one of the renowned news bulletins that includes

worldwide reports, expert opinions, public opinions, features on interesting topics and current affairs. This program is very popular among the public. These video bulletins were converted into two audio files of wav format. Each audio file consisted of three different news stories. From the first audio file which was in Aalia Nazki's voice, a detailed news about an exhibition scheduled to be held in Mohatta palace was selected. And from the second audio file of Shafi Taqi Jami, the main headlines of all three news were selected. These three headlines were further broken into three audio files. Each audio file played a news headline.

In order to play these news streams a different information design strategy i.e. multiple information communication simultaneously was used. In this bulletin, the detailed exhibition news was set to play continuously throughout the bulletin in a female voice. This voice stream was termed as a primary voice in the experiment. Moreover, while keeping the primary voice in playing mode the other three news stored in three audio files were also played after periodic intervals of 10 seconds. This voice considered as a secondary voice. The primary voice was set to come from left earphone whereas the secondary voice was set to come from right earphone. This approach was adopted because it was expected that playing primary and secondary voice in different earphones would bring ease for the user to discriminate both voice streams.

These two files with given information design were merged into one clip and played by writing a program in Visual Studio 2013 using C#. The total duration of this clip was 1 minute and 28 seconds. This clip was played on Dell Vostro 5560 with Core i5 processor and 4GB RAM. In order to listen to the clip, an average quality KHM MX earphone was used to listen to the clip.

The experiment was conducted on people ranging from 20 to 55 years including both males and females. Total 10 users participated in this experiment out of which 6 were male and 4 were female. The experiment was conducted at random places without considering whether the environment / surrounding was fully quiet or not.

In order to judge the behavior of users in the experiment, a questionnaire was prepared. The interviewees were first briefed about the audio playing mechanism in this experiment. They were told about both the primary and the secondary voices. Before they started to listen to the audio clip they were given an overview of the questionnaire so that they could grab the information accordingly. The questionnaire aimed to establish whether a listener could notice, focus and comprehend multiple information simultaneously or not. It also helped to gauge the notice, selection and attention behavior of the user. In order to facilitate and reduce the memory load, participants were given maximum three choices to select one from.

5.2 Results

Most of the users were able to answer the questions correctly which were asked to find out, whether they could hear both the sounds simultaneously or not. And when they were asked about the perceptual and observational question all of the participants found voice streams audible, discriminable when played together.

Following is the response of users for each question asked in the questionnaire.

i. Could you hear the primary voice presenting documentary? From all participants, 80% of the users told that the primary voice presenting documentary was clear.

The remaining 20% users who although said that they were able to listen to the primary voice but remarked that it was loud and shrilling so could further be improved.

ii. What was the topic of primary voice? All the participants rightly told the topic of primary voice i.e. Exhibition.

iii. Where was the exhibition scheduled to hold? Fifty percents of the users could not answer it correctly. Remaining those who answered correct, guessed it using their prior knowledge. The use of user's existing knowledge behavior would fully be investigated in upcoming series of experiments.

iv. What was the venue name? All the participants correctly answered the venue name of exhibition i.e. 'Mohatta Palace'.

v. Could you please tell us more about the exhibition documentary? In order to judge users' comprehension, they were asked to describe what they listened in the exhibition documentary. All the users were able to describe the documentary and gave the overview in broken words. These words were kind of keywords in the documentary that users used.

It is observed that though users lost some amount of information while focusing on secondary voice but they still grasped the documentary very well and where there was an information gap they filled it with their existing knowledge.

vi. Could you notice the secondary voice? Yes, all users were able to notice the secondary voice in the presence of primary voice.

vii. Were you able to distinguish secondary voice in the presence of primary voice and vice versa? 70% users stated that they found no difficulty to distinguish secondary sound from primary voice and vice versa. The 30% users were of the view that it could further be improved.

It is learned that this easiness in discriminating both the voice streams was mainly possible, because, both sounds were coming in different ears separately and also the voice streams were uttered by different gender voices i.e. male and female. In order to make 'discrimination' more evident, other auditory dimensions could also be explored.

vii. Were you able to distinguish secondary voice in the presence of primary voice and vice versa? The 70% users stated that they found no difficulty to distinguish secondary sound from primary voice and vice versa. Remaining 30% users were of the view that it could further be improved because they missed some information while focusing a particular voice.

It is learned that this easiness in discriminating both the voice streams was mainly possible, because, both sounds were coming in different ears separately and also the voice streams were uttered by different gender voices i.e. male and female. In order the make 'discrimination' more evident, other auditory dimensions could be explored which we would do in future experiments.

viii. What was secondary voice indicating? All the users correctly answered that secondary voice was indicating news.

ix. How many times secondary voice played in different intervals? The 30% users gave a wrong answer while 70% users rightly told that it was played three times.

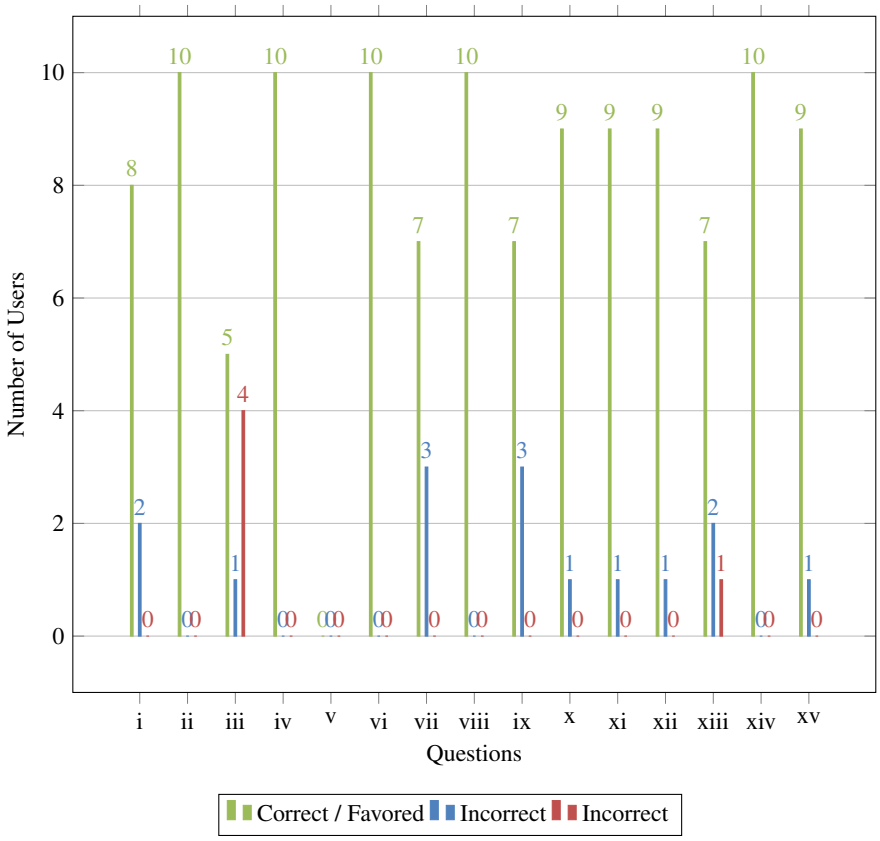

Fig. 2. Users' response performance in multiple voice-based information communication

The above bar-chart indicates the number of correct / incorrect answers by the users for each question. The question v is discriptive, therefore, not reflected in bars whereas bars in question xiii indicate the selection of interested news by the users from three headlines played to them. In question i and vii the second blue bar indicate that how many users had asked to improve the quality.

x. In the first occurrence, what was the topic of secondary voice? Among all participants, only one user couldn't answer this question correctly.

xi. In the second occurrence, what was the topic of secondary voice? The 90% of participants correctly told the topic of the second occurrence i.e. cyber attack.

xii. In the third occurrence, what was the topic of secondary voice? Same results were witnessed as seen in above two questions.

xiii. Which was the most interesting news for you? The 70% users opted 'Data theft in Cyber Attack' remaining twenty percent of the users opted for 'Black Money in Budget' whereas only one female user showed interest in Exhibition Documentary.

xiv. Did you want to promptly listen to the detail news from any of the spoken news? As a follow-up to Question 13 when users asked to tell their intent that whether they wanted to promptly listen to the interesting news by skipping the present primary voice then 100% of the users answered 'yes'.

This is an interesting finding which provides the opportunity of applying GUI based overlay, lightbox techniques in voice-based interaction which is discussed in the previous section using the Facebook wall of a user.

xv. Did you find multiple sounds helpful in reaching multiple information quickly and Would you prefer this approach over the sequential flow of information? The 90% of the users found this quick design of delivering information helpful and said they would prefer this multiple information communication simultaneously over the sequential flow of information. From these 90% users, a few had reservations. They said, in this technique they are afraid that they might loose some important information which they would prefer to listen without any noise and disturbance. So, it could also be an interesting finding that in which contexts the multiple information communication design strategy could be applied and where it can't.

The 10% of the users who didn't give preference said they are uni-task oriented so can't prefer this approach over the sequential flow of information.

6 Conclusion and Future Work

The results of this experiment are encouraging to further explore this design approach. The results validate that multiple information communication is possible using voice in Human-machine interaction. Users showed interest in multiple information communication. They were able to discriminate the voice. Using their focus and attention abilities they were able to get multiple information meaningfully in lesser time.

We find it suitable to further investigate this information design approach. We are presently in the process to develop a software that would be able to play multiple live programs simultaneously. Each program would have its own set of controls mapped with auditory perceptions. Users would be able to set the controls, i.e. they would be able to pan the stream, make the volume low and high, change the pitch, change the rate of voice streams and much more which may help them to listen to multiple voice streams simultaneously using their focus and attention abilities. This web-based software would be used to observe the interaction behaviour of users. For example, what values they set to the control to listen to the multiple sounds?

References

1. Guo-ping Li and Guo-yong Huang. The "core-periphery" pattern of the globalization of electronic commerce. In *Proceedings of the 7th International Conference on Electronic Commerce*, ICEC '05, pages 66–69, New York, NY, USA, 2005. ACM.
2. Raman Kazhamiakin, Piergiorgio Bertoli, Massimo Paolucci, Marco Pistore, and Matthias Wagner. Having services âĂIJyourway!âĂİ: towards user-centric composition of mobile services. In *Future Internet–FIS 2008*, pages 94–106. Springer, 2009.

3. Philip Kortum. *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.
4. Dirk Schnelle-Walka. I tell you something. In *Proceedings of the 16th European Conference on Pattern Languages of Programs*, page 10. ACM, 2012.
5. World Health Organization. Visual impairment and blindness. `http://www.who.int/mediacentre/factsheets/fs282/en/`, 2014. [Online; accessed 04-Jan-2016].
6. Ádám Csapó and György Wersényi. Overview of auditory representations in human-machine interfaces. *ACM Computing Surveys (CSUR)*, 46(2):19, 2013.
7. Alan Dix, Janet E Finlay, Gregory D Abowd, and Russell Beale. Human-computer interaction. 2003.
8. Daisuke Sato, Shaojian Zhu, Masatomo Kobayashi, Hironobu Takagi, and Chieko Asakawa. Sasayaki: Augmented voice web browsing experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2769–2778, New York, NY, USA, 2011. ACM.
9. Karen Church, Mauro Cherubini, and Nuria Oliver. A large-scale study of daily information needs captured in situ. *ACM Trans. Comput.-Hum. Interact.*, 21(2):10:1–10:46, February 2014.
10. Sheetal K. Agarwal, Anupam Jain, Arun Kumar, Amit A. Nanavati, and Nitendra Rajput. The spoken web: A web for the underprivileged. *SIGWEB Newsl.*, (Summer):1:1–1:9, June 2010.
11. MPuerto Paule-Ruiz, Víctor Álvarez García, J. R. Pérez-Pérez, and M. Riestra-González. Voice interactive learning: A framework and evaluation. In *Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE '13, pages 34–39, New York, NY, USA, 2013. ACM.
12. Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S. Parikh. Avaaj otalo: A field study of an interactive voice forum for small agriculturers in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 733–742, New York, NY, USA, 2010. ACM.
13. James R Lewis. *Practical speech user interface design*. CRC Press, Inc., 2010.
14. Ronald L Schow, J Anthony Seikel, Gail D Chermak, and Matthew Berent. Central auditory processes and test measuresasha 1996 revisited. *American Journal of Audiology*, 9(2):63–68, 2000.
15. Roxanne L Canosa. Real-world vision: Selective perception and task. *ACM Transactions on Applied Perception (TAP)*, 6(2):11, 2009.
16. João Guerreiro. Using simultaneous audio sources to speed-up blind people's web scanning. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, page 8. ACM, 2013.
17. Ibrar Hussain, Ling Chen, Hamid Turab Mirza, Gencai Chen, and Saeed-Ul Hassan. Right mix of speech and non-speech: hybrid auditory feedback in mobility assistance of the visually impaired. *Universal Access in the Information Society*, pages 1–10, 2014.
18. João Guerreiro and Daniel Gonçalves. Text-to-speeches: evaluating the perception of concurrent speech by blind people. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 169–176. ACM, 2014.
19. Michael W Eysenck. *Psychology: An international perspective*. Taylor & Francis, 2004.
20. Bill Scott and Theresa Neil. *Designing web interfaces: Principles and patterns for rich interactions*. " O'Reilly Media, Inc.", 2009.