USING SPEECH TO SEARCH: COMPARING BUILT-IN AND AMBIENT SPEECH SEARCH IN TERMS OF PRIVACY AND USER EXPERIENCE

R. Bernhaupt, D. Drouet, F. Manciet, M. Pirker and G. Pottier

Institut de Recherche en Informatique de Toulouse (IRIT), France

ABSTRACT

While voice assistants are on the rise for a variety of applications, talking to the television still feels less natural to users than talking to their friends or neighbours. Subtle differences how and when a voice assistant is activated seem to influence the overall perception of users in terms of user experience, control and acceptance of such an interactive system. To investigate the influence of using speech to search for content with an ambient voice assistant, compared to a more traditional solution with a microphone in a remote control, an experimental study was performed. Fourteen participants took part in a within-subject experiment comparing an ambient speech interaction with speech search using a remote control with a dedicated button to activate the microphone in terms of: privacy, usability and user experience. Results indicate a slightly higher impression of control for the button-based speech search modality, as well as fewer privacy concerns by the users for the button-based speech search modality. In terms of user experience, the hands-free ambient speech search does not perform significantly better than the traditional button-based speech search approach.

INTRODUCTION

The usage of speech interaction is on the rise: labelled by marketers as "voice assistants"' a variety of products has appeared on the market. In this paper we use *speech* to refer to a user talking to a system, while *voice* is used to refer to a person's unique voice that allows identification.

Since the introduction of “Siri” on Apple’s iOS on mobile devices back in 2011, voice assistants have gained popularity and get smarter every day. All major brands presented voice assistants: Cortana (Microsoft), Google Assistant (Google) (1), Alexa (Amazon) (2). With Alexa, Amazon started a new trend: now the voice assistant is ambient, it is not only embedded in the user’s smartphone or laptop, but has its own dedicated device, which is standing in the home, always listening and at the user's service (3). While Alexa was the only device designed to be a standalone vocal assistant back in 2014, Google released Google Home and “Spot” (4), the connected ambient microphone by nvidia which will allow the user to control her home and her nvidia Shield TV in a hands-free manner, thanks to the Google Assistant embedded in the device. Microsoft’s Cortana will also be embedded in a speaker device, designed by Harman Kardon.

All these ambient vocal assistants are changing the speech-search landscape. When it comes to TV, until now the user had to use the built-in microphone on the TV or set-top-box remote control to perform her voice search. Now, with the ambient microphone, users will be free to perform their speech search without having to talk to their remote control, and will have a complete hands-free experience. This leads us to these questions: are users ready to open their homes to these technologies? What are the differences between this “ambient” modality and a voice-controlled remote control solution, in terms of usability, user experience, feeling of control, acceptance, and privacy concerns for the TV use case? Is there still room in the living-room for the remote control?

The goal of this study was to understand if the ambient speech search is enhancing the user experience of watching TV, compared to a traditional voice search using the remote control’s built-in microphone. We also wanted to measure how determining factors, such as acceptance of such a technology and user’s privacy concerns, are influencing the overall experience of the user.

STATE OF THE ART

The use of speech as a means of interaction with a system comes from the area of natural language processing. Elder (5) is considered to be the first one who uses speech as an interaction technique. Speech for interaction has the advantage that people can use their natural way to talk to the system as they would talk to others. Unfortunately, technological systems introduce a certain number of technological concepts (like volume) that can lead to a much more complicated interaction when using natural language. Instead of pushing a button several times for volume up, the user has to either use Volume up, up, up, up, up (5) or refer to labels they might not be familiar with (set volume to 40 might be a very loud setting for the TV).

Overall, speech is a convenient modality to assist users when searching for content. Discovery activities can be supported as well as browsing large collections of content (7,8, 9). In a comparative study where the users had to browse long lists of channels using a remote control or their voice, Ibrahim et al (7) found out that users preferred at the time to have a TV remote control, in which the microphone is embedded, to freely talk to their TV, rather than talking directly to their TV screen.

In another experiment, Stifelman et al (10) designed a speech-based search system that allows users to freely talk to their TV without the need of any remote control. Reactivity of the speech-to-text system and visual hints about what the system understood were highlighted as key design guidelines to provide a pleasant and reassuring experience. Nonetheless, they did not investigate how users would react to the fact that their TV is always listening to them, nor how would users specify to the TV that they are talking to it. The use of such ambient listening devices raises new challenges for speech interaction, not only in terms of usability and user experience but especially for privacy.

Privacy concerns and trust in technology are now challenges for tech companies in general, but even more when it comes to the household environment. Sailaja et. al. (11) investigated privacy within a user study about electronic program guides (EPG). They found that users were interested in getting a personalized EPG. However, they were expressing a deficit of trust about the treatment of their personal data. The lack of

transparency from the service providers was an issue for them. As this aspect of trust is crucial for such sensitive data as voice, we also wanted to investigate this point in our study.

TALKING TO THE TV: AN EXPERIMENT

Research Questions and Method

The goal of this experiment is to understand the difference between an ambient voice assistant and a voice assistant activated with a button press on a remote control in term of usability, user experience, acceptance, and privacy.

Our hypotheses were:

- (H1) If the vocal assistant is used with the ambient approach, user experience will be rated higher compared to the remote control approach.
- (H2) If the vocal assistant is used with the remote control approach, the acceptance will be higher compared to the ambient approach;
- (H3) If the vocal assistant is used with the remote approach, then the perceived usability will be higher than if it was used with the ambient approach;
- (H4) If the vocal assistant is used with the ambient approach, users will have more privacy concerns than if it was used with the button-based modality.

To investigate the above hypotheses an experiment was performed: each participant performed three sets of content discovery tasks on the same TV system (see Figure 1). In one set of tasks, participants used the ambient voice search interaction, while in another set the remote control button-based interaction was used. After using both interactions, participants were free to choose the interaction they preferred, or to combine both during the third set of tasks. User experience in terms of hedonic and pragmatic quality as well as attractiveness was measured with the Attrakdiff questionnaire, meaning and value as UX dimensions were investigated in the final interview questions.

Acceptance was judged following the user acceptance of hedonic systems (12). Based on three pillars of the Technology Acceptance Model (13): perceived ease of use; perceived usefulness; and behavioural intention of use extended with perceived enjoyment. According to Davis, 1989 (14), perceived ease of use refers to “the degree to which a person believes that using a particular system would be free of effort”. Perceived usefulness is “degree to which a person believes that using a particular system will enhance his or her performance”. Behavioural intention of use indicates whether or not the user wants to use the system. And finally, the perceived enjoyment refers to “the extent to which the activity of using the system is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated” according to Davis et. al. (14). These four dimensions are measured in a questionnaire by Van Der Heijden (12).
Privacy concerns were measured in interviews with the users at the end of the study, after they had tried the ambient and the remote control approaches. Users were additionally asked about their general concerns about privacy, using online services, their concerns about using voice assistant for TV, and if they feel a difference between the two approaches.

The feeling of control was assessed during an interview, after users tried both approaches. Usability was measured using 5-scale questions for naturalness (1 being very natural, 5 being not natural at all) and difficulty (1 being very easy, 5 being very difficult) at the end of each task.

Prototype, Participants and Procedure

For this experiment a prototypical interactive TV system (programmed in Unity3D) with functions like live TV, electronic program guide (EPG), video on demand (VoD) or weather information was used. A voice search engine is embedded in the prototype allowing the user to be very specific in the way she is asking her queries. Natural language is supported thanks to the API wit.ai and the built-in speech-to-text engine that comes with Unity3D. The system also keeps the last query in its memory to offer the possibility to refine the results obtained from the immediate previous query. For example, it is possible to ask “I would like to watch a comedy” which will give a list of 49 comedies. The user can then refine the query asking “only the recent ones”, which will reduce the number of comedies shown to 21, to show the comedies released during the past five years. Naturally, if the user asks for the “recent comedies” directly, she will obtain immediately the same results. Various types of filters are available, amongst them actors, directors, genres, scores, studios (e.g. Marvel, Pixar), type of content (e.g. content based on true events, biopic), title (e.g. Titanic, The Terminator), channels, period of the day (e.g. this evening, tomorrow, this afternoon).

Search by speech interaction can be used to find VoD content or live TV information in an easy way within the user interface. Comparing two pieces of content is also possible, giving an indication which content is the more popular and has the better score. Based on these features, a set of four tasks has been developed. Each task requires the user to combine several filters using speech interaction before selecting and buying / watching content.

Slight variations of these four tasks must be achieved with both approaches. For the remote interaction, the user has to press and hold the microphone button, and the voice search menu will be displayed. The TV immediately shows what was understood by the system in this menu when the user is talking (see Figure 1). For the ambient modality, a catchphrase needs to be stated to trigger the voice search menu, such as “Hey TV” or “Hey Alison”, and then the users can formulate their queries.

Figure 1 - On the left, remote control used during the survey, Linemote. On the right, voice search menu of our user interface, listening state.

Fourteen users participated in this experiment, seven males and seven females, aged 19 to 25 (M = 21) mostly students at the university. One participant reported to use speech interaction on a regular basis, the others did not have prior experiences with speech-based user interfaces before taking part in the study. They were all regularly watching TV (several times a week).

The experimental procedure was as follows: (1) welcoming participants; (2) filling out camera allowance and consent form; (3) trial task to get acquainted with the system; (4) four tasks with speech interaction to perform tasks (either ambient or remote control) followed by questionnaires; (5) four task with an alternative speech interaction system followed by questionnaires; (6) four tasks with the real speech interaction system followed by (7) questionnaires; and interview questions. After each task, users answered questions about the perceived difficulty of the task, the naturalness, as well as meaning and value of the interaction modality.

Sessions were video-recorded with two cameras. One was in front of the user, to focus on users' faces, and one was above the users, to check how they were interacting with the remote control. Studies were conducted in French and participants received 20 EUR for their participation in the study.

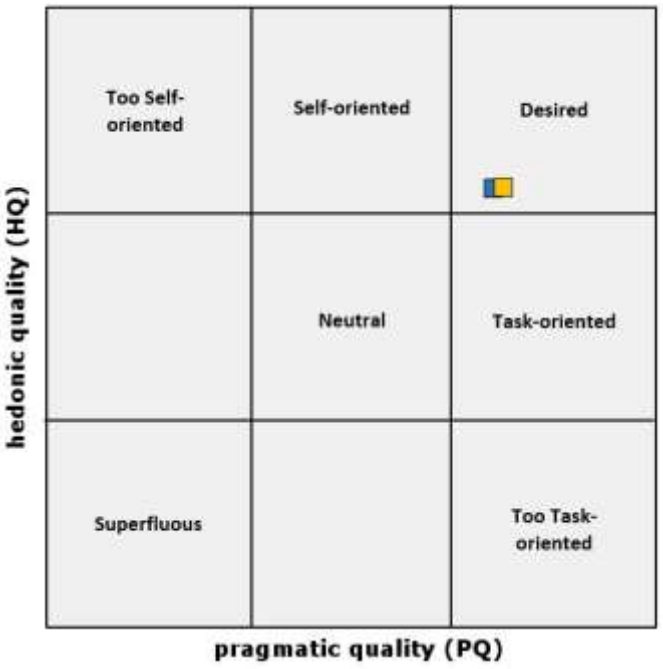

Figure 2. AttrakDiff results. Blue: Ambient modality. Orange: Button-based modality.

RESULTS

(H1) User Experience is higher with the ambient approach: not validated

As Figure 2 indicates, user experience was rated very similar for the two conditions. Interview questions at the end of the study investigated participants' preferences for the two modalities. Speech as a mechanism for search was well appreciated: 13 participants preferred to use their voice for content discovery tasks compared to menu-based interactions with French IPTV system which do not embed speech as interaction mechanism in their search engines. The use of speech for these tasks appeared to be very meaningful to the users.

Additionally, the value rating obtained after each task did not reveal significant differences between the two modalities (see Table 1). Although users claimed that they do appreciate the use of speech search during their content discovery activities, it remains unclear if they prefer the ambient solution over the remote control solution, or if users simply do not care.

After the use of the two interactions, we asked users for their favourite. Nine users preferred the RCU interaction, while five of them preferred the ambient one. There were various reasons for their choices. For the ambient interaction, it was the simplicity aspect

(hand free, effortless and simple), while for the remote it was efficiency (faster, no catchphrase, control in hand). One user perceived the remote interaction as less awkward than the ambient, and another one choose the remote because she was uncomfortable with the ubiquity of the ambient microphone.

Tasks	Task 1		Task2		Task 3		Task 4		Overall (Mean)	
Modality	AMB	RCU	AMB	RCU	AMB	RCU	AMB	RCU	AMB	RCU
Mean	1.21	1.21	1.79	1.50	1.50	1.50	1.57	1.57	1.51	1.44
SD	0.42	0.42	1.05	0.85	0.94	0.76	1.08	0.75	0.55	0.44
T	1.50		3.00		5.00		8.50		17.00	
P	1.000		0.214		1.000		0.666		0.509	

Table 1. Value ratings (1: very valuable; 5: not valuable at all) for all the four tasks with the two modalities (AMB: ambient, RCU: button-based) showing no significant differences in rating (students t-test with t value and probability p)

When it comes to the feeling of freedom, the ambient interaction was preferred by nine participants compared to the five who felt more freedom using the RCU interaction. The two most stated reasons were that with the ambient interaction that hands are free for doing other things (4 times) and that you do not have to search for the remote control (3 times).

	Perceived Usefulness		Perceived Ease of Use		Perceived Enjoyment		Behavioural Intention of Use	
Modality	AMB	RCU	AMB	RCU	AMB	RCU	AMB	RCU
Mean	2.18	2.14	1.69	2.14	1.94	2.01	1.10	1.07
SD	0.62	0.55	1.33	1.17	0.82	0.90	1.07	1.10
t(13)	0.425				-0.380		0.123	
T			17.50					
P	0.678		0.090		0.710		0.904	

Table 2. Acceptance questionnaire results (AMB: ambient, RCU: button-based).

(H2) Acceptance is higher with the button-based approach: not validated

Results from the acceptance questionnaire did not give clear indications of which solution would be better in terms of overall (see Table 2). In terms of perceived usefulness, perceived ease of use, perceived enjoyment and behavioural intention of use, there were no significant differences. Additionally, the intention of use, measured on a 10-point scale (1 being “I really want to use it”, 10 being “I don’t want to use it at all”) and rated after the user had performed each set of tasks with each system, do not show significant differences (ambient, M=3.07, SD=1.94; button-based M=2.5, SD=0.855; p=0.25).

Concerning which interaction they will use at home if they have the opportunity, three of the participants choose the ambient solution, five of them choose the RCU, while six stated that they would use both. This was explained by referring to the context of usage: if the remote control is out of reach, it's more convenient to talk directly to the TV because it does not force the user to move from the couch. On the other hand, if the remote control is already in the user's hand it's easier and more natural to talk using the remote control's built-in microphone.

(H3) Perceived usability is higher with the button-based approach: partially validated

No significant difference between the two interaction modalities has been observed in terms of perceived difficulty for each task (see Table 3). On the contrary, naturalness ratings expressed after each task are significantly different for the two modalities; the button-based modality was evaluated as more natural than the ambient modality. This finding was confirmed in the final interview where users were asked which interaction they perceived as the most natural. Nine users stated it was the button-based solution, while four stated it was the ambient one (one user did not feel any difference). Amongst the reason for their choices, users highlighted again the fact that the button-based modality does not require catchphrases. The remote control is a solution users are already accustomed to use, and it is perceived as a very natural interaction. Additionally, ten participants stated that they felt more in control with the remote-based approach, compared to two participants for the ambient approach, and another two participants that did not perceive any differences between the two voice modalities.

Tasks		Task 1		Task2		Task 3		Task 4		Overall (Mean)	
Perceived difficulty	Modality	AMB	RCU	AMB	RCU	AMB	RCU	AMB	RCU	AMB	RCU
	Mean	1.29	1.21	1.36	1.71	1.43	1.07	1.21	1.43	1.32	1.35
	SD	0.61	0.57	0.63	1.20	0.93	0.26	0.57	0.64	0.45	0.34
	T	4.00		4.00		0.00		4.50		32.00	
	p	0.705		0.163		0.102		0.408		0.577	
Naturalness	Mean	2.07	1.79	2.43	2.00	2.43	1.50	2.21	1.71	2.21	1.75
	SD	1.54	1.12	1.22	1.17	1.35	0.76	1.42	0.99	1.27	0.78
	T	6.00		19.00		5.00		9.50		11.00	
	p	0.339		0.379		0.121		0.222		0.026	

Table 3. Difficulty rating (1: very easy; 5: very difficult) and naturalness (1: very natural; 5: not natural at all - AMB: ambient, RCU: button-based).

(H4) Privacy concerns are higher with ambient modality: confirmed.

After using both interactions we asked users about their privacy concerns in general using online services (social networks, video streaming, shopping services). Twelve users expressed privacy concerns, although these concerns did vary greatly between the users

– it ranged from being careful about bank data to using tape to obscure the web cam of their personal computers.

After this general question, we explained the differences, in terms of data collection, between the ambient and the remote control approaches. Then we asked users if they have some privacy concerns about voice interaction in general. Eleven users claimed to have privacy concerns with voice, and three did not care. They were afraid of being spied on by the companies which are delivering the content (9 users out of 11), and would prefer that all the voice records stay on a local device, under their control. Finally, we asked users if they perceive differences in terms of privacy protection between the two modalities and 10 of them stated that the ambient microphone was less reliable. Again, it is the fear of being spied that motivated their answer. The button-based modality gave them the feeling they can control what the device records and when it records their voice. The 4 remaining users did not perceive any differences. The hypothesis that users felt more in control with the remote control was thus confirmed.

SUMMARY AND DISCUSSION

Results of this small-scale user study shed new light on the usage of voice interaction in the living room. Using voice is extremely convenient when it comes to content discovery activities. It supports rich interaction and in-depth searches, thanks to natural language processing. It is valuable to the user and it has a great potential to provide a playful experience with conversational user interfaces. Yet, participants of our study mentioned that talking to the TV is still awkward and not adapted to social situations. These situations still represent 60% of the time spent in front of the TV (15). Overall the design of voice interactions should better consider contextual aspects.

Our comparison of ambient voice search and button-based voice search revealed that the user's opinion on, and experiences with, the system are not affected by the modality being used. Contrary to our expectations, also the user experience did not perform better for the ambient voice search modality, as we would have expected due to novelty and innovativeness of this approach. Nonetheless, the high number of privacy concerns expressed about the ambient modality might be a barrier in the adoption and use of such systems. The TV remote control remains a reliable device over which the user believes she has a full control. This control is particularly appreciated when it comes to voice, as voice represents the user in a unique manner, and it carries information about user lifestyle and opinions, which are sensitive.

Voice remains the future modality of the living room, but still has to overcome limitations and social barriers. A recent survey from Tivo (16) suggests that in the United States and Canada, only 8.5% of people use their voice to find something to watch, and within this 8.5%, only 15.3% of users use this feature on regular basis. Such a low score is consistent with our findings. Voice is valuable, but it is still unnatural to talk to the TV, especially if the TV gives the feeling that it is spying on you. These insights suggest that there is still room for improvement to make this feature an unmissable hit in the living room.

REFERENCES

1. Google Assistant (retrieved 18.5.17). https://assistant.google.com/
2. Amazon Alexa (retrieved 18.5.17). https://www.amazon.com/Amazon-Echo-BluetoothSpeaker-with-WiFi-Alexa/dp/B00X4WHP5E
3. Google Home (retrieved 18.5.17). https://madeby.google.com/home/
4. Nvidia Spot (retrieved 18.5.17). https://www.nvidia.com/en-us/shield/smart-home/
5. Elder, H. A., 1970. On the feasibility of voice input to an on-line computer processing system. Communications of the ACM, June, 1970. pp. 339 to 346.
6. Igarashi, T., and Hughes, J. F., 2001. Voice as sound: using non-verbal voice input for interactive control. Proceedings of the 14th annual ACM symposium on User interface software and technology. November, 2001. pp. 155 to 156.
7. Ibrahim, A., Lundberg, J., and Johansson, J., 2001. Speech enhanced remote control for media terminal. INTERSPEECH. 2001. pp. 2685 to 2688.
8. Ibrahim A. and Johansson P., 2002. Multimodal Dialogue Systems for Interactive TVApplications. Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI '02). October, 2002. pp. 117.
9. Wittenburg, K., Lanning T., Schwenke, D., Shubin, H., and Vetro A. 2006. The prospects for unrestricted speech input for TV content search. Proceedings of the working conference on Advanced visual interfaces (AVI '06). May, 2006. pp. 352 to 359.
10. Stifelman, L., Elman, A., and Sullivan, A., 2013. Designing natural speech interactions for the living room. CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13). April, 2013. pp. 1215 to 1220.
11. Sailaja, N., Crabtree, A., and Stenton, P., 2017. Challenges of using personal data to drive personalised electronic programme guides. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. May, 2017. pp. 5226 to 5231.
12. Van Der Heijden, H., 2004. User acceptance of hedonic information systems. MIS Quarterly, 28. 2004. pp 695 to 705.
13. Davis, F. D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly 13, 3. September, 1989. pp. 319 to 340.
14. Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. 1992. Extrinsic and Intrinsic Motivation to Use Computers in the Workplace. Journal of Applied Social Psychology (22:14), 1992. pp. 1111 to 1132.
15. Vanattenhoven, J. and Geerts, D., 2015. Contextual aspects of typical viewing situations: a new perspective for recommending television and video content. Personal and Ubiquitous Computing. Vol. 19, no 5-6. 2015. pp. 761 to 779.
16. TiVo. 2016. Q4 2016 Video Trends Report: Consumer Behavior Across Pay-TV, VOD, PPV, OTT, TVE, Connected Devices, and Content Discovery.