“Read That Article”: Exploring Synergies between Gaze and Speech Interaction

Diogo Vieira[1,2], João Dinis Freitas[2], Cengiz Acartürk[3], António Teixeira[1], Luis Sousa[2], Samuel Silva[1], Sara Candeias[2], Miguel Sales Dias[2,4]

DETI / IEETA
University of Aveiro
Campus Univ. de Santiago
Aveiro - Portugal
{diogo.vieira, ajst, sss}@ua.pt

Microsoft Language Development Center
Lisboa, Portugal
{t-joaof, t-luisno, midias, t-sacand}@microsoft.com

Cognitive Science
Middle East Technical Univ
Ankara, Turkey
acarturk@metu.edu.tr

ISCTE-IUL
ISCTE - University Institute of Lisbon
miguel.dias@iscte.pt

ABSTRACT

Gaze information has the potential to benefit Human-Computer Interaction (HCI) tasks, particularly when combined with speech. Gaze can improve our understanding of the user intention, as a secondary input modality, or it can be used as the main input modality by users with some level of permanent or temporary impairments. In this paper we describe a multimodal HCI system prototype which supports speech, gaze and the combination of both. The system has been developed for Active Assisted Living scenarios.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Input devices and strategies, Natural language, Voice I/O*; K.4.2 [**Computers And Society**]: Social Issues - *Assistive technologies for persons with disabilities;*

Keywords

Multimodal; Gaze; Speech; Fusion

1. INTRODUCTION

A major goal of Human-Computer Interaction (HCI) research, is to achieve natural forms of interaction, similarly to the way humans interact. Speech is, in that sense, a dominant modality since it is usually the preferred way for humans to communicate.

Nevertheless, speech communication is inherently a multimodal process where different senses are used for interpreting spoken messages. Previous studies, notably the one that discovered the McGurk effect [5], revealed that humans employ both hearing and visual senses in speech perception. Humans also use contextual information such as head and body movements, gestures, facial expressions, characteristics of the speech signal like prosody and gaze in human-human communication [3,6,7].

In this context, we propose that the integration of user’s gaze information (for instance, the fixation of one or both eyes on the computer screen), has the potential to improve the determination of the intention of the user, during a natural interaction task. Gaze information can help to predict the user purpose and, when used with speech in command and control scenarios, to decrease the confusion of the issued commands, thus improving the accuracy rates of the speech recognition systems [4, 8]. A recent review can be found in Acartürk et al. [1]. Given the potential benefit of processing user’s gaze location, we decided to add this modality to an existing system designed for elderly users in the context of several Ambient Assisted Living (AAL) initiatives [2].

2. GAZE MODALITY

A new modality was developed and added to the existing W3C based multimodal framework developed by the authors, for projects such as the Portuguese QREN AAL4ALL and European AAL PaeLife [2, 9]. In this architecture the creation and integration of new modules is simplified, since it supports the requirements of a multimodal framework, such as extensibility, flexibility and the use of multiple input or output modalities. Some of these, already built for the framework, are touch, speech or gesture as input modalities, and text to speech synthesis as output modality. In our framework, every user interaction results in a message that is sent to an interaction manager responsible for handling all the communication and fusion across all the modalities.

This new system component provides not only the possibility for the user to interact directly with an application using only gaze, but also use it in combination with the other input modalities, such as speech. Thus, also regarding the targeted user base, two problems were simultaneously tackled: using gaze as a single modality brings the possibility for the group of users with certain permanent or temporary disabilities to interact and use the application; when simultaneously used with another input modality, fusing both interactions allows us to obtain a multimodal user experience were we are capable of deriving the intention of the user, in a less ambiguous way and with more confidence. Gaze as a modality uses an eye tracker device to obtain the user’s gaze input data in real time, by means of fixations and saccade events. In parallel, the application keeps updating the modality context with the current list of 2D graphical objects painted on the application screen, which correspond to application actions that can be triggered by gaze fixations. When our modality component detects a gaze fixation over a given graphical object, as a consequence of the user staring at such object on the screen, the modality sends the matching unimodal or multimodal input command associated with that object to the interaction manager.

3. FUSION WITH SPEECH

The system currently supports speech interaction in European Portuguese, English, French, Hungarian and Polish. Spanish and Turkish will be available soon. Although several languages are supported, using speech as the only input interaction modality may suffer from some limitations including, for example, the low

ASSETS '15, October 26-28, 2015, Lisbon, Portugal.
ACM 978-1-4503-3400-6/15/10.
http://dx.doi.org/10.1145/2700648.2811369

level of confidence of the utterance of the user obtained from the recognizer, either due to word mispronunciation, co-articulation errors, ambient noise or precision limitations of the ASR.

To improve the recognition of the intention of the user when communicating with the system using speech, one possibility is to use information regarding the application context, narrowing the set of expected user input commands. The use of gaze can provide us with that needed context. The fusion of gaze and speech brings us the possibility of using the gaze information to disambiguate through the list of recognized words and phrases and their likelihood, obtained from the ASR process of the speech modality. This fusion process might follow two different paths: both inputs modalities result on the same output, or they don't. In the first case, the user looks at an object on the screen, and at the same time pronounces the voice command to trigger the action assigned to such object. When producing these actions, two equal commands are sent from different input modalities processors, to the same interaction manager. Since there are no differences in both commands, the fusion only calculates a new confidence score, issuing a single command to the application, which corresponds to our best guess of the intention of the user. In the second case, the fusion process interprets each input command for completeness and confidence score calculation and takes in consideration the differences in such interpretation, with the objective of merging their information together to achieve a more complete multimodal input command. That is possible by taking the incomplete unimodal command (for example, speech) and adding information obtained from the other complete unimodal command (for instance, gaze).

4. EXAMPLES OF USE

In this paper, we've added the support to gaze, as well as to speech+gaze acting as single input multimodality. With our technique, when the user's fixation is directed to a button on the PLA interface with some time persistence, it will issue a command from the gaze modality processor that triggers the same event as if the button was clicked, e.g., after 5 seconds (configurable parameter) of a user fixation on the news reader button, the application will open that option. A list of news will be then presented on the application screen and each one can also be opened using only gaze fixation, much like in the prior example. Additionally, any news article can be open if the user pronounces its title name or the first words in the news. We have further improved this user experience, with our novel multimodal approach. Eliciting his/her intention by using gaze and speech simultaneously, the user may simply utter 'read that article' while at the same time looking at the desired news title he/she wants to read, in the app screen. And although the speech command does not explicitly specify the wanted news title, our gaze modality can provide the missing information on the object the user was looking at.

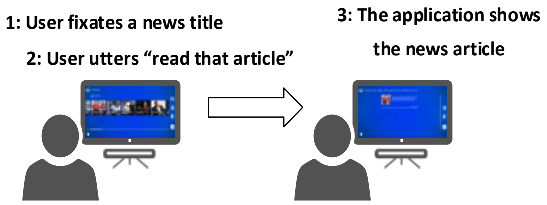
1: User fixates a news title
2: User utters "read that article"
3: The application shows
the news article

Figure 1. The application navigates successfully when the user's gaze is used to complement a speech command

5. CONCLUSIONS AND FUTURE WORK

Exploring how speech and gaze can be used together in the context of a multimodal user experience with an app such as the PaeLife PLA, is an important first step towards a more natural user experience. This work describes several use scenarios and is the first to our knowledge to combine, within a multimodal framework, speech recognition in European Portuguese with gaze, serving as a ground for further research. The adoption of gaze as a new modality creates multiple possible ways to interact with applications, developed within our multimodal framework. Also, the use of speech and gaze input modalities for tasks which are inherently performed with the eyes, e.g., visualization in Virtual or Augmented Reality, and for which speech can be used to perform tasks such as zooming or picking, with gaze providing the place-of-interest, is an interesting route to further explore in the near future. This aspect was already captured by the pioneering work of Richard Bolt from the MIT, back in 1980, in his case, by fusing speech and gesture in a visualization task in a media room. By taking advantage on our modular multimodal framework, and adding other modalities in the fusion process, the possible multimodal ways for interacting with future applications are large. Given the current state-of-the-art in the combined use of speech and gaze for HCI, there are still several challenges ahead, (e.g. eye-tracking calibration, environmental noise, etc.), particularly if we consider the development of mission critical applications. Our aim for future work is to develop a stochastic-based fusion model based on the probability density functions modelled for both gaze and speech. To achieve this goal we need not only to be able to model the user intent (for a given task), but also to tackle deeply issues such as signal synchronization and delays between input events of different modalities.

6. ACKNOWLEDGMENTS

This work was fully funded by Marie Curie IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP) http://www.iris-interaction.eu/

7. REFERENCES

[1] Acartürk, C., et al. Elderly speech-gaze interaction: State of the art and challenges for interaction design. *Proc. HCII 2015*, (2015), (to appear).

[2] Almeida, N. and Teixeira, A. Enhanced interaction for the Elderly supported by the W3C Multimodal Architecture. *Proc. Conf. Nacional sobre Interacção*, (2013).

[3] Hakkani-Tür, D., et al. Eye Gaze for Spoken Language Understanding in Multi-modal Conversational Interactions. *Conf. on Multimodal Interaction*, ACM (2014), 263–266.

[4] Heck, L.P., et al. Multi-Modal Conversational Search and Browse. *SLAM@ INTERSPEECH*, (2013), 96–101.

[5] McGurk, H. and MacDonald, J.Hearing lips and seeing voices. *Nature 264*, (1976), 746–748.

[6] Oviatt, S.Ten myths of multimodal interaction. *Commun. ACM 42*, 11 (1999), 74–81.

[7] Quek, F., et al.Multimodal human discourse: gesture and speech. *ACM Trans. on Comp.-Human Interaction (TOCHI) 9*, 3 (2002), 171–193.

[8] Slaney, M., et al.Gaze-enhanced speech recognition. *Proc. ICASSP*, (2014), 3236–3240.

[9] Teixeira, A., et al. Speech-centric Multimodal Interaction for Easy-to-access Online Services-A Personal Life Assistant for the Elderly. *Procedia Computer Science 27*, (2014), 389–397.

[10] Bolt,. Richard, "Put-that-there, Voice and Gesture at the Graphics Interface", A*CM Proceedings SIGGRAH* (1980)