

Vocate: Auditory Interfaces for Location-based Services

John McGee
Digital Media Centre (DMC)
Dublin Institute of Technology (DIT)
Dublin, Ireland
Tel: 00353 1 402 3270
johnnyboymcgee@gmail.com

Charlie Cullen
Digital Media Centre (DMC)
Dublin Institute of Technology (DIT)
Dublin, Ireland
Tel: 00353 1 402 3273
Charlie.cullen@dit.ie

ABSTRACT

This paper discusses work being carried out by the *Vocate* module of the *LOK8* project. The *LOK8* project seeks to develop location-based services within intelligent social environments, such as museums, art galleries, office buildings, and so on. It seeks to do this using a wide range of media and devices employing multiple modalities. The *Vocate* module is responsible for the auditory aspect of the *LOK8* environment and will seek to exploit the natural strengths afforded by the auditory modality to make the *LOK8* system user-friendly in multiple scenarios, including instances where the user needs to be hands-free or eyes-free, or when screen size on a mobile device might be an issue. We look at what kinds of services the *Vocate* module will be seeking to implement within the *LOK8* environment and discuss the strengths and weaknesses of three possible approaches - sonification, auditory user interfaces, and speech interfaces.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces – Auditory (non-speech) feedback, Evaluation/methodology, Interaction styles (e.g., commands, menus, forms, direct manipulation), User-centered design; H.5.2 [User Interfaces (D.2.2, H.1.2, I.3.6)]: Natural language, voice I/O; I.2.7 [Natural Language Processing].

General Terms

Design, Human Factors.

Keywords

Auditory user interfaces, sonification, speech interfaces, location-based services, contextual awareness, audio navigation.

1. INTRODUCTION

The *LOK8* project's objective is to deliver context-specific, location-based services within an intelligent environment. It seeks to do this using a wide range of media in multiple modalities via screens,

projectors, head-mounted displays (HMDs), mobile devices, speakers, and so on. The *LOK8* system will make use of media within the environment to provide scaleable content depending on the context, location and personal preferences of the user. In its most immersive form the *LOK8* environment will present users with personalised, interactive avatars that will guide them via speech and gestural interaction but beyond this it will seek to exploit the advantages afforded by multiple modalities to make content delivery scaleable and to make the *LOK8* environment user-friendly in situations where it might not be practical or desirable to attend to a visual display or manual interface. The project is divided into four distinct modules: the *Vocate* module, which handles the auditory aspect of the environment; the *Avatar* module, which handles the visual aspect of the environment; the *Tracker* module, which handles positioning and locationing within the environment; and the *Contact* module, which provides the dialog system for the environment. This paper details work within the *Vocate* module of the project relating to auditory interfaces. It discusses the advantages that audition has over other modalities and outlines what types of services *Vocate* will be trying to implement using audio. It will also consider how *Vocate* might realise these implementations using three possible approaches: sonification, auditory user interfaces and speech interfaces.

2. VOCATE

The *Vocate* module will seek to implement a number of features within the *LOK8* environment. Firstly, it will seek to provide a hands-free navigation system that can both guide users to target destinations within the environment as and when they are requested, and also point out salient information relating to the environment itself (or objects within the environment) as and when it becomes relevant to the user's spatial context. This type of navigation system reduces the necessity for visual aids such as maps, which can be cognitively demanding in situations where you are in transit and may need to focus on your immediate surroundings; they can also be impractical on mobile devices where screen real estate might be at a premium – a key issue given the trend towards smaller and smaller handsets in many modern mobile devices. Secondly, *Vocate* will seek to provide an auditory version of the *LOK8* environment's menu interface that can be interacted with remotely, either via the user's mobile device or possibly via intercoms located throughout the environment. Such a menu system would allow users to continue to interact with the *LOK8* environment even in situations where their focus and attention cannot be devoted to the manual operation of their mobile device; it would also remove the burden on the visual modality when screen space on a mobile device is limited. Finally, *Vocate* will seek to provide realistic speech interaction with the *LOK8* avatar when it is in operation within the environment. This multimodal approach

in particular will seek to provide the most immersive and natural interactive experience within the *LOK8* environment and will likely be collaborative across all four project modules. It is the aim of the *LOK8* project that this multimodal, avatar-based approach will lead to more intuitive, naturalistic human-computer interaction, and away from the physically constrained, traditional methods of computer interaction such as the mouse, keyboard, and even the touchscreen. Each of these design tasks has its own unique challenges, *Vocate* will be considering these in relation to three possible approaches: sonification, auditory user interfaces, and speech interfaces.

3. AUDITORY INTERFACES

Audio information is processed faster neurally than both haptic and visual information (2ms for audio, compared with 10ms for haptic and 100ms for visual information [13]), this lends the auditory modality well to the delivery of certain types of information, such as alerts and alarms, particularly when one considers that audio notifications are generally harder to ignore than visual notifications. Audio is also hands-free and largely focus-independent, which makes it a suitable modality for the delivery of information in scenarios where the user may be in transit or have their eyes and/or hands occupied with a cognitively demanding task [11]. Factors such as these, combined with the fact that ever-improving technology is allowing for acceptable quality audio to be increasingly possible on smaller and cheaper devices, place audio in a unique position when considering multimodal solutions to user interface design problems and physicality issues.

3.1 Sonification

Sonification is defined as the use of non-speech audio to convey information [14]. The underlying concept of sonification has been around for many years, early examples would include the hourly chimes of a clock tower to convey the time of day, the foghorn, and Morse code. Today modern technology allows designers to incorporate sonification systems into a wide range of devices. There are, however, a number of inherent obstacles when it comes to sonification that the sound designer must consider. Firstly, not all types of information are suitable for sonification. For example it may be quite straightforward to get a listener's attention by using a high-frequency alert but what if the designer then wants to use sonification to communicate something quite complex to the listener, such as the identity of people who work in the building they are currently in? This brings us to our second obstacle - lack of established design conventions. While there are numerous examples of systems that have sonified quite complex information and data sets, such as pie charts [9], daily weather records [8], market information [12], and patterns in DNA and RNA sequences [6], many would argue that the field of sonification still lacks established design conventions. Sound design does not have the same wealth of recognised guidelines and design principles that the visual arts have, perhaps because audio is less tangible in nature, but organisations such as ICAD (International Community for Auditory Display) and ISCRAM (Information Systems for Crisis Response and Management) are working to change this. A third factor the designer must consider is the environment in which the sonification system is to be used. We make use of a considerable amount of auditory information in our surrounding environment on a daily basis, this would include naturally occurring sounds as well as existing auditory displays, such as doorbells and telephone ringtones. One must take care to design a sonification system that can work in tandem with this ambient information and not against it, this can be done effectively

by studying the target environment and testing any proposed systems in comparable conditions [1][17].

3.1.1 Relevance of Sonification to *Vocate*

In terms of the *Vocate* project sonification could be of particular use when it comes to the implementation of the audio navigation system. Sonification lends itself well to the communication of spatial information within an environment because the information being conveyed is generally physical in nature rather than abstract and therefore simpler to convey. The use of stereo spatialisation and volume modulation can allow the sound designer to 'place' auditory information within the soundscape as if it were coming from an actual physical location relative to the user. This approach has been used in several systems to communicate the location of both target destinations and objects of interest within an environment [18][20][24]. While some systems, such as the *Ontrack* system [24], use the listener's music of choice as the source signal for spatialisation and modulation, there is also existing empirical research into the efficacy of using beacons that could be leveraged for the *Vocate* project [20][23]. The beacon approach generally uses spatialised sonar style pulses of sound to indicate a destination or path through an environment, the tempo or volume of the beacon signal usually increases as the user approaches the target destination. Over longer distances several beacons may be placed between the user and the target destination, as each beacon is reached the next one in the series becomes active. Studies have found that broad spectrum sounds, such as pink noise bursts, are more easily localised and have been found to encourage greater performance. It has also been found that a moderate capture radius (i.e the area within which the system deems the user to have reached the beacon, thus triggering the next beacon to become audible) is preferable to a very large or very small capture area e.g. greater than 9ft or only a few inches.

On a slightly more abstract level sonification has also been used to communicate when a user has moved from one surface or area to another [23], for example moving from the pavement onto the road, or from the Italian Renaissance section of an art gallery to the Romanesque section. What makes this more abstract is that with this approach these different surfaces and areas have to be allocated their own unique acoustic characteristics in some way so as to differentiate them from each other and there aren't always natural acoustic mappings available to the designer, the level of difficulty in this regard depends on the context and in some cases it may be more expedient to use speech notifications.

The *LOK8* project aims to be used in social settings such as museums, art galleries and shopping centres, with this in mind it must be considered that using headphones or earphones could discourage social interaction between users within the environment because each user would be operating within their own private audio space. Previous sonification systems have attempted to address this issue, for example Stahl's *Roaring Navigator* [18] developed an 'eavesdropping' system whereby if multiple users were in close physical proximity, those who were not currently listening to anything in their headsets could pick up a certain amount of the audio that other users were listening to. Another possible solution to this issue would be the use of bonephones. Bonephones are open-ear headphones that use vibrations to transmit sound directly to the cochlea via the bones of the skull thus allowing external ambient audio to remain audible via the ear canals. Tests have shown that although bonephones do not perform as well as headphones when

it comes to stereo spatialisation they are still sufficiently effective when used in audio navigation scenarios [23]. The unique physical advantage of being able to bypass the outer ear completely also allows bonephones to be used by anyone suffering from conductive hearing loss.

3.2 Auditory User Interfaces

An auditory display is defined as the use of sound to communicate information about the state of an application or computing device to a user; this definition suggests the unidirectional flow of information from the device to the user. An auditory user interface on the other hand is defined as a superclass of auditory displays that allows for auditory input to also flow from the user back to the device, usually in the form of speech [15]. By this nature auditory user interfaces are less constricted than sonification or speech interfaces alone and as such are the easiest to integrate into a multimodal environment. In the past a lot of research in the field of auditory user interfaces has been driven by the need to develop alternative user interfaces for the visually impaired but it has since come to be seen as an area of considerable potential in its own right, both in terms of exclusively auditory user interfaces and augmented audio-visual user interfaces, such as the *JMusic* system [4], which allows users to map the runtime behaviours of Java programs onto musical parameters and hence monitor these behaviours continually. The ability of audio to operate on the periphery of a user's awareness is particularly useful in this regard as it can allow a system or device to give continual feedback without necessarily leading to cognitive overload. Many mechanical devices physically generate sounds during operation that over time users learn to interpret as indicative of the operational status of the device as a whole e.g. the way in which a mechanic might listen to an engine to hear what's wrong [11]. The digital nature of many modern devices has, in many cases, done away with this physical form of feedback but a carefully considered auditory ecology within any system can reintroduce some of this functionality.

3.2.1 Relevance of Auditory User Interfaces to *Vocate*

In terms of the *Vocate* project auditory user interfaces arguably offer the best option for the implementation of the auditory version of the *LOK8* menu interface. The *LOK8* menu interface will be the most basic mode of interaction with the *LOK8* system, offering access to all of the functionality that the *LOK8* environment has to offer; this might include the ability to query objects of interest within the environment, the ability to query one's location within the environment, guided tours within the environment, information relating to available services and amenities, and so on. While the *LOK8* menu will also likely feature a traditional graphical user interface, a stand-alone auditory version of the menu will offer equivalent functionality in situations where the user requires to be hands-free and/or eyes-free, or when screen space on a mobile device is limited - the main advantage of non visual user interfaces in terms of physicality is that they effectively render the physical issue of screen size redundant.

Audio is serial in nature and while this offers some advantage over the visual modality when it comes to complex data comparisons [5] [16], it is something of a weakness when it comes to menu design as the visual modality, unlike its auditory counterpart, can quite easily continually present multiple objects of interest, such as menu options, to the user. Despite this physical limitation there are still inherent qualities in audio that lend themselves to menu design. The human auditory system is particularly adept at filtering audio

information into perceptually meaningful elements by a process that Bregman describes as 'auditory scene analysis' [3]. The three main aspects of auditory scene analysis are segregation, segmentation and integration. The human auditory system applies these filtering techniques to divide audio information into 'streams'; a stream might be made up of one quick audio event such as a loud bang (an example of segregation), or it might be made up of a collection of associated sounds such as a choir singing (an example of integration). Whether sounds are segregated, segmented or integrated with other sounds depends on several parameters including pitch, frequency, timbre, volume, tempo, spatial location, and so on. An example where the human auditory system uses these phenomena to great advantage is 'the cocktail party effect' [2], whereby a listener can zone in on one speaker in a room full of conversations and extraneous noise. Empirical studies regarding the parameters and thresholds that effect auditory stream perception have enabled sound designers to design auditory systems that can present users with multiple streams of audio information in such a way that the streams can be kept perceptually separate from each other and brought in and out of focus when necessary [7][10]. This, combined with techniques such as 'skimming' (the ability to skim segments of an audio stream to give an indication of the whole stream), help counteract some of the negative aspects of audio seriality. *Vocate's* auditory user interface could adopt a combination of this speech-based approach along with other techniques and principles leveraged from sonification to give the user additional feedback regarding the operation of the menu system.

3.3 Speech Interfaces

Speech interfaces are interfaces that utilise speech recognition and/or speech synthesis to communicate with a user. The obvious advantage that speech interfaces have over other auditory interfaces, such as sonification systems, is that they can communicate with the user using natural language. Having said that, speech interfaces are arguably the most difficult and time-consuming of all auditory interfaces to implement and speech itself brings with it its own problems. An obvious problem with adopting a speech interface is that language becomes a factor. While sonification can often transcend linguistic and cultural boundaries speech interfaces are limited to the languages that the system and its users share common knowledge of. While speech is often the best option for communicating highly complex or specific information it is not necessarily a suitable option when communicating ambient information; it is often common in ambient displays to abstract the data being communicated in some way in order to render the display easier to interpret and experience on a peripheral level [19]. Speech interfaces also require a lot of back-end work. The corpora with which the system is trained have to be rigorously compiled and the most effective speech interface systems use multiple forms of data input, such as lip-tracking, gaze-tracking, and gestural input, in order to model the system's responses and output. This is because speech communication is generally quite physical in nature, with much information and meaning conveyed via body language and backchannel communication; failure to address this physical aspect of speech communication can lead to less efficient speech interface systems.

3.3.1 Relevance of Speech Interfaces to *Vocate*

We have already discussed how speech interaction might be highly suitable for aspects of the *LOK8* auditory menu interface but it will also have application when users are interacting with the *LOK8* avatar. The goal with the avatar is to have a character that the

user can interact with as if it were a real personal assistant or tour guide. The fact that all four modules of the *LOK8* project will be collaborating on this aspect of the environment in particular means that the multimodal approach necessary to achieve effective human-computer speech interaction should be possible. For example the *Vocate* module will look at speech signal processing along with the *Contact* module, which will also be working on the dialog manager and modeling how the avatar will behave and react in relation to the input the *LOK8* system receives. The *Avatar* module will not only be working on the visual design and aesthetic of the avatar, but will also be looking at optical recognition for the purposes of obtaining gestural input. Finally, the *Tracker* module will allow the *LOK8* system to display the audio and visual output in the correct context for the user based on their location and position. It is the goal of the *LOK8* team that a well-rendered avatar-based interface with audio-visual input and output will encourage more natural and intuitive interaction between the user and the environment and transcend some of the physical constraints of more traditional human-computer interaction methods.

A further option in relation to speech interfaces is that there are now several off-the-shelf products available on multiple platforms and mobile devices, such as *Vlingo* (available on Blackberry, iPhone, Nokia and Windows Mobile), *Voice Control* (Apple's new speech interface system for the iPhone 3GS), and *Google Mobile App* (available on iPhone), that promise a lot of the functionality that *LOK8* is seeking to implement. Although testing would be required on such products to ensure that they both possess the requisite functionality, and adequately plug in to the rest of the *LOK8* system, they would certainly be an option worth considering.

4. CONCLUSIONS

In this paper we outlined the work of the *LOK8* Project and specifically the role the *Vocate* module plays within that project in developing auditory interfaces. We discussed some of the unique qualities that audition has to offer as a modality for communication and interaction, such as its hands-and-eyes-free nature and fast neural processing rate. We discussed the pros and cons that different auditory interfaces might offer in terms of the specific services *Vocate* seeks to implement within the *LOK8* environment i.e. a hands-and-eyes-free navigation system, an auditory menu interface, and natural speech interaction with an avatar. Sonification lends itself well to the communication of physical information and any form of information that has natural acoustic mappings but it is not always suitable for complex, detailed interactions. Speech interfaces can be highly effective when it comes to communicating directly with a user in more complex interactions but they require a lot of back-end work as well as multiple forms of data input for more naturalistic systems. One must also consider that language may become an obstacle when using speech interfaces as a certain level of fluency with the language(s) used by the system may be required of the user, unlike with sonification which can often transcend such linguistic boundaries. Auditory user interfaces offer umbrella solutions that can leverage strengths from both sonification and speech interfaces but one must take care to consider the environment the auditory system is to be deployed in and make use of existing empirical data wherever possible. Finally we discussed the fact that recent off-the-shelf products offer much of the functionality that the *LOK8* system seeks to offer and that such devices might be worth considering should they stand up to testing within the overall *LOK8* environment.

5. REFERENCES

- [1] Alexanderson, P. 2004. Peripheral Awareness and Smooth Notification: The Use Of Natural Sounds In Process Control Work. In Proceedings of the 3rd Nordic Conference on Human-computer Interaction (Tampere, Finland, October 23 - 27, 2004). CHI'04. ACM. 281-284.
- [2] Arons, B. 1992. A Review of the Cocktail Party Effect. Journal of the American Voice I/O Society. 12 (July, 1992). 35-50.
- [3] Bregman, A. S. 1999. Auditory Scene Analysis: The Perceptual Organization Of Sound. The MIT Press, London.
- [4] Collberg, C., Kobourov, S., Hutcheson, C., Trimble, J., and Stepp, M. 2005. Monitoring Java Programs Using Music. Technical Report TR05-04. Department of Computer Science, University Of Arizona.
- [5] Cullen, C. and Coyle, E. 2004. Analysis of Data Sets Using Trio Sonification. In Proceedings of the Irish Signals and Systems Conference (Belfast, Northern Ireland, June 30 - July 2, 2004). ISSC'04.
- [6] Cullen, C. and Coyle, E. 2003. Rhythmic Parsing of Sonified DNA and RNA Sequences. In Proceedings of the Irish Signals and Systems Conference (Limerick, Ireland, July 1 - 2, 2003). ISSC'03.
- [7] Fernström, M. and McNamara, C. 2005. After Direct Manipulation - Direct Sonification. ACM Transactions on Applied Perception (TAP). 2 (October, 2005). 495-499.
- [8] Flowers, J. H., Whitwer, L. E., Grafel, D. C., and Kotan C. A. 2001. Sonification Of Daily Weather Records: Issues Of Perception, Attention And Memory In Design Choices. In Proceedings of the International Conference on Auditory Display (Espoo, Finland, July 29 - August 1, 2001). ICAD'01. 222-226.
- [9] Franklin, K. M. and Roberts, J. C. 2003. Pie Chart Sonification. In Proceedings of 7th International Conference on Information Visualization (July 16 - 19, 2003). IV'03. IEEE Computer Society. 4-9.
- [10] Frauenberger, C. and Stockman, T. 2006. Patterns In Auditory Menu Design. In Proceedings of the 12th International Conference on Auditory Display (London, UK, June 20 - 23, 2006). ICAD'06. 141-147.
- [11] Gaver, W. W., Smith, R. B. and O'Shea, T. 1991. Effective Sounds In Complex Systems: The Arkola Simulation. In Proceedings of the Conference on Human Factors in Computing Systems: Reaching Through Technology (New Orleans, Louisiana, April 27 - May 02, 1991). ACM New York. SIGCHI. 85-90.
- [12] Janata, P. and Childs, E. 2004. Marketbuzz: Sonification Of Real-time Financial Data. In Proceedings of the 10th Meeting of the International Conference on Auditory Display (Sydney, Australia, July 6 - 9). ICAD'04.
- [13] Kail R. and Salthouse, T.A. 1994. Processing Speed As A Mental Capacity. Acta Psychologica. 86, 2-3 (June, 1994). 199-255.
- [14] Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N., and Neuhoff, J. 1999. Sonification Report: Status of the Field and Research Agenda. Technical Report. ICAD, 1999.
- [15] McGookin, D. 2004. Understanding and Improving the Identification of Concurrently Presented Earcons. PhD thesis, University of Glasgow. 155-159.
- [16] Potard, G. 2006. Guernica 2006: Sonification of 2000 Years of War and World Population Data. In Proceedings of the 12th International Conference on Auditory Display (London, UK,

- June 20 - 23, 2006). ICAD'06.
- [17] Sanderson, P. M., Shek, V., and Watson, M. 2004. The Effect Of Music On Monitoring A Simulated Anaesthetised Patient With Sonification. In Conference of the Human Factors and Ergonomics Society of Australia Special Interest Group in Computer Human Interaction (Wollongong, Australia, November 28 - December 1, 2004). OzCHI'04.
- [18] Stahl, C. 2007. The Roaring Navigator: A Group Guide For The Zoo With Shared Auditory Landmark Display. In Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services (Singapore, September 9 - 12, 2007). Mobile HCI 07. ACM. 383-386.
- [19] Terry, M. A., Tran, Q. T., and Mankoff, J. 2001. Ambient Displays: A Designer's Synopsis. In Proceedings of CHI'01 Extended Abstracts on Human Factors in Computing Systems (Seattle, Washington, March 31 - April 05, 2001). ACM Press.
- [20] Walker, B. N. and Lindsay, J. 2006. Navigation Performance With a Virtual Auditory Display: Effects of Beacon Sound, Capture Radius, and Practice. *Human Factors*. 48, 2, (Summer 2006). Human Factors and Ergonomics Society. 265-278.
- [21] Walker, B. N., Raymond, S. M., Nandini, I., Simpson, B. D., and Brungart, D. S. 2005. Evaluation Of Bone-Conduction Headsets For Use In Multitalker Communication Environments. In Proceedings of the Human Factors And Ergonomics Society 49th Annual Meeting (Orlando, Florida, September 26 - 30, 2005). Human Factors and Ergonomics Society. HFES'05. 1615-1619.
- [22] Walker, B. N. and Raymond, S. M. 2005. Thresholds Of Audibility For Bone-Conduction Headsets. In Proceedings of the 11th Meeting of the International Conference on Auditory Display (Limerick, Ireland, July 6 - 9, 2005). ICAD'05. 218-222.
- [23] Walker, B. N. and Lindsay, J. 2005. Navigation Performance In A Virtual Environment With Bonephones. In Proceedings of the 11th Meeting of the International Conference on Auditory Display (Limerick, Ireland, July 6 - 9, 2005). ICAD'05. 260-263.
- [24] Warren, N., Jones, M., Jones, S., and Bainbridge, D. 2005. Navigation Via Continuously Adapted Music. (Portland, Oregon, USA, April 2 - 7, 2005). CHI'05. ACM. 1849-1852.