HARD AND SOFT INFORMATION FUSION IN SONIFICATION FOR ASSISTIVE MOBILE DEVICE TECHNOLOGY

Jeff Rimland and David Hall
College of Information Sciences and Technology,
The Pennsylvania State University,
University Park, PA 16802
jrimland@ist.psu.edu, dhall@ist.psu.edu

Mark Ballora
College of Arts and Architecture,
The Pennsylvania State University,
University Park, PA 16802
ballora@psu.edu

ABSTRACT

Current research integrating sonification with hard and soft information fusion is poised to enable a new class of assistive technology that is driven by sensor technology and informed by semantic understanding of the user's capabilities and current goals as well as events unfolding in the world around them. At the time of this writing, Penn State University is in the fourth year of an Army Research Office (ARO) funded MURI grant to develop networked hard and soft information fusion techniques with the main goal of using "soft data" (including textual and spoken reports from military personnel; and weather, news, and nearby hazard reports from the web and publicly available social media feeds) along with "hard data" from a variety of physical sensors (including video, acoustic, LIDAR, and rangefinder) with the goal of detecting and preventing the use of improvised explosive devices (IEDs) against deployed troops. Over the course of this research, we have identified several areas in which the sonification-based utilization of hard and soft information can offer significant improvements to existing assistive technologies and tools. This paper presents the key elements of this research, describes areas of applicability in assistive technology, and presents suggestions for future research.

1. INTRODUCTION

Most technological tools for assisting humans to perform a task or better understand an environment rely on a specific modality of data or information. Under ideal conditions, humans are able to quickly (and often subconsciously) gather a sufficiently comprehensive mental model of their situation via the use of such single-modality tools in conjunction with their own sensory information, memory of past situations, and ability to reason about future outcomes. However, when any of these information sources or human capabilities become impaired through either environmental factors (e.g. smoke obscuring vision) or biological causes (e.g. visual or cognitive impairment), the difficulty and time expenditure of obtaining information and mentally fusing it can result in a condition where it is infeasible or impossible to maintain situational awareness.

Consider a person preparing to leave their house to walk across town for an appointment. While brushing their teeth they might refer to a calendar in order to confirm the specific time and location of the appointment, quickly check news and weather using a smart phone or web browser, and then glance outside to ensure that the reported conditions are accurate and that no other unforeseen conditions or hazards exist. That person then proceeds toward their appointment with the full capability of all of their senses ready to alert them to emerging threats or changing conditions.

Using existing tools such as text-to-speech technology, a visually impaired person could also access the same information prior to leaving the house. However, it would be difficult for them to continue to monitor those information sources while navigating toward their appointment. Changes in weather, traffic accidents, and any number of natural or man-made emergencies could arise. In essence, the same individuals who are most vulnerable to unexpected dangers are least equipped to obtain alerts to those dangers in real time.

A similar situation can exist for deployed military, emergency, or fire fighting personnel who do not have a disability. Responding to dangerous visually intensive situations while monitoring situational awareness tools for unseen and emerging threat information can quickly lead to a condition of sensory and cognitive overload.

In both cases described above, the ideal assistive technology would automatically monitor potentially useful "soft" information sources while also relying on physical sensors to augment shortcomings of the user's own sensory system. In the military case, this physical sensor might be a laser-based system such as LIDAR [1] that can provide "super-human" ability to detect threats from long distances through darkness and dense smoke. For assisting visually impaired users, the sensor contribution might be utilization of the built in camera on a mobile device to provide auditory cues related to navigation, threat detection, and object recognition.

In either case, the challenge is to automatically merge and present real-time multi-source, multi-modality information to a user who is constantly on the verge of information overload in a manner that offers an effective cognitive and sensory prosthesis without inducing undue distraction or discomfort. Sonification is a key aspect of this approach for users who are either visually impaired, working in environments of limited visibility, or performing critical and visually-intensive tasks that must not be interrupted by referencing additional visual displays.

2. AUGMENTING CURRENT ASSISTIVE TECHNOLOGY

One goal of this paper is to introduce a new sonification technique with broad applicability to improving existing assistive tools. For providing context on applying these improvements, we will focus on the Georgia Institute of Technology's System for Wearable Audio Navigation (SWAN) [2], and also refer to other assistive and sonification technologies as appropriate. The SWAN system currently provides: 1) a Navigation Beacon, which guides the user along a pre-determined path via a series of waypoints; 2) Object Sounds, which announce the location and type of nearby objects detected; 3) Surface Transitions, which indicate that a walking surface is about to change (from grass to sidewalk, for example); 4) Locations such as classrooms and restaurants; and 5) Annotations, which are spoken messages left by other users who encountered difficulties at that location previously (e.g. deep puddles form here after a heavy rain") [3]. The following sections will briefly describe how each of these sonification challenges can be improved or augmented through the use of hard/soft information fusion.

2.1. Navigation

To enable audio-based navigation in SWAN, Georgia Institute of Technology researchers have developed an effective probability-based localization method [4], explored the human factors of using auditory alerts for navigation [5], and even developed a Virtual Reality (VR) environment for testing and prototyping [6]. However, detecting dynamic changes to the environment through physical sensors alone has inherent limitations in reacting to unforeseen events only as they reach the effective detection range of the hardware. Integration with open source news, weather, and social media reports via software agents [7] and determining potential threats along a future route (both ahead of time and as the trip progresses) can reduce risk without requiring the user to monitor multiple news sources manually before leaving the house and also while navigating. Additional examples of such sensor-based electronic travel aid (ETA) systems that could potentially benefit from integration with "soft" data include the Naviton prototype [8] developed at Lodz University of Technology and the AudioGuider system [9] developed at Zhejiang University.

In [10], Heuten et al present a sonification framework for blind or visually impaired users to virtually explore an unfamiliar city prior to actually visiting that location. They suggest the following methods for avoiding information overload when using auditory objects to represent Geographic Information System (GIS) entities:

1) Filter Objects
2) Clustering Objects
3) Level of Detail Modulation
4) Object Prioritization via Volume
5) Distance-based Prioritization
6) Object size-based Prioritization
7) Zooming and Panning

These approaches to reducing the number of sound sources (and thereby reducing information overload) can also apply to systems that utilize both hard and soft information sources for real-time navigation guidance. For instance, instead of prioritizing objects solely based on their distance or size, they could be evaluated based on both physical attributes and relevant semantic data extracted from news sources or Internet search data.

2.2. Object Detection

Reliable sensor-based object detection is very difficult to accomplish without the additional guidance of contextual information about the task that the user is currently trying to accomplish. Integration with calendar/task management tools can reduce the dimensionality of the object search space by more heavily weighting detection results that are consistent with objects likely to be encountered during the current task.

Calendar and scheduling assistance has been demonstrated to be well suited to the multi-agent software paradigm (see [11], [12], [13]), but there has been limited research in extending calendar agents to provide assistance with other tasks such as object detection that could benefit from the increased context that could be provided by converting scheduling information into semantic data formats that support inferencing and sense-making across multiple data modalities.

Semantically informed object detection is particularly useful for visually impaired individuals receiving sonified representations of objects in crowded or rapidly changing environments. Contextually-aware software agents in such an environment could both improve the success rate of the object identification process (by reducing the size of the search space to objects likely to occur in the current situation), and reduce information overload by only "displaying" objects likely to be relevant. For example, if the system knows that a blind user is preparing a meal, it could use a single tone to represent non-cooking items, and more varied tones to help the user determine which container holds salt and which holds pepper. If the user is vacuuming, then salt and pepper containers might be represented by the same tone, because their auditory representation is merely intended as background information to communicate location of possible obstacles, as opposed to foreground items that are central to the current task currently being attended to.

2.3. Surface Transitions and Locations

Although some location and surface information can be readily obtained by cameras/sensors alone (such as the pedestrian crossings referred to in [14]), most information of this nature must be obtained from maps or other databases.

The Geographic Information Systems (GIS) community and the Open Geospatial Consortium (OGC) have developed a wide variety of tools, standards, and protocols for representing layers of highly varied information ranging from air quality, traffic and weather data to emergency response and disaster management notifications. Much of this data is openly

available to the public. However, the challenge lies in: 1) querying the correct data sources with the parameters that will return relevant data to the user, 2) converting the vastly heterogeneous data into a common format that allows sense-making across multiple data modalities, and 3) presenting the results to the user in a manner that improves their situational awareness without causing undue distractions.

The robotics community has utilized a variety of Synchronous Location And Mapping (SLAM) techniques (such as in [15]) to build maps and feature databases in real time as robots are exploring their surroundings. Similar techniques could be applied to supplying attributes of walkways and other location data (railings, stairs, handles, etc.) to visually impaired people navigating challenging environments. See [16] for an example of specific image processing techniques applied to detection of walkway surfaces intended to be safer for the visually impaired.

2.4. Annotations

Allowing users to record audible messages and warnings for other users who find themselves in the same location can provide valuable information. However, this can be vastly enhanced by the inclusion of massive publicly available sources of geocoded and annotated imagery and video from services such as Twitter, Flickr, and YouTube. Again, the challenge lies in effectively querying this data and optimally providing it to the users, which is addressed by the information architecture shown in Figure 1.

Additionally, this type of functionality introduces the challenge of working with auditory displays that are at the intersection of *Auditory Information Displays* and *Sonification*. According to the sonification design space map introduced in [17], auditory information displays provide well-defined sound cues such as synthesized speech and alert tones, while sonification renders data into sound for the purpose of human analysis and perception. Auditory annotations based on fused hard and soft data would deliver a combination of auditory information display (in the form of spoken messages) as well as sonification of aggregated data. For example, a dangerous area or condition might be mentioned hundreds or thousands of times on twitter. Listening to the actual text of thousands of messages in real time is not feasible, but listening to a sonification of the aggregated sentiment analysis [18] of those messages might only take a few seconds. The user would then have the option of listening to a sample of the spoken text at their discretion.

3. TECHNIQUES FOR HARD AND SOFT INFORMATION FUSION

The key challenges in this area include: 1) Feature extraction; 2) Data encoding, representation, and storage; 3) Fusion of multi-source, multi-modality data; 4) User interactions; and 5) Test and evaluation. This section will briefly describe each of these challenges, summarize key research findings, and provide references for further exploration.

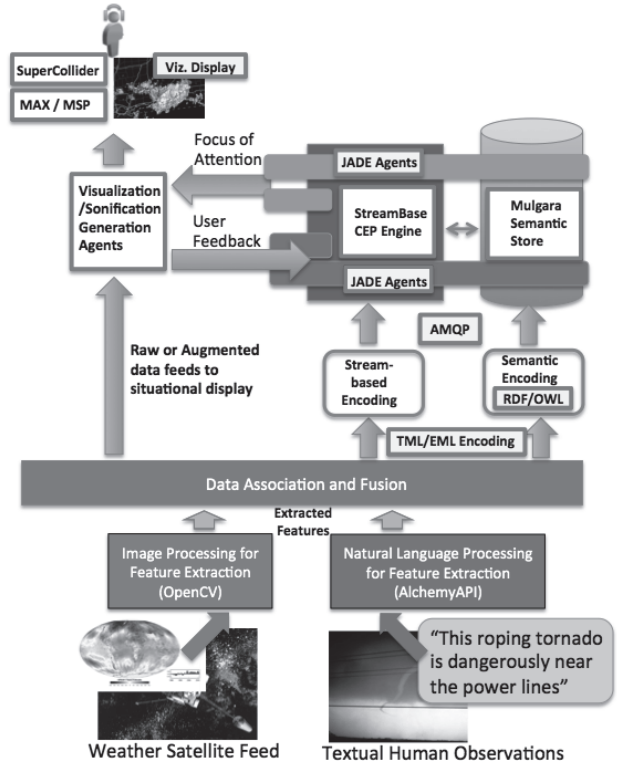

Figure 1: An extensible architecture for sonification of fused hard and soft information from distributed sources (adapted from [18]).

3.1. Feature Extraction

Many excellent open source tools are available for low-level feature extraction. These include OpenCV for image processing [19] and AlchemyAPI for extraction of relevant warnings and sentiment analysis of open data sources [20]. Any software architecture intended to utilize open source data can benefit greatly from embracing the web service paradigm for feature extraction. Otherwise, large amounts of time can be wasted reinventing functionality that has already been implemented by the open source community and made available via web API endpoints.

Many examples of low-level feature extraction in the context of hard and soft information fusion are available in [21].

3.2. Encoding, Representation, and Storage

In this research, Open Geospatial Consortium (OGC) guidelines for representation of geospatial, sensor, and human-centric data and observations were followed [22]. Transducer Markup Language (TML) and Event Pattern Markup Language (EML) [23] were critical for handling heterogeneous data as well as interacting with existing tools and web services. Resource Description Framework (RDF) triples and semantic stores such as Mulgara [24] were also used to represent feature vectors of both hard and soft information.

Additionally, Open Sound Control (OSC) was selected both for its interoperability with tools such as SuperCollider and its efficient low-latency communication over User Datagram Protocol (UDP) [25].

3.3. Fusion of Multi-Source, Multi-Modality Data

One contribution of particular interest to the sonification and accessibility community is the novel combination of Complex Event Processing (CEP) and multi-agent systems (MAS) to enable integration of top-down and bottom-up information processing flows, human-in-the-loop interaction via sonification, and optimized dynamic access to third-party web services and information sources. This new paradigm also facilitates corroboration of data from hard and soft sources using dynamic temporal, geospatial, and complex event windows.

The Complex Event Processing (CEP) paradigm is well suited to high speed processing of data from multiple input streams. It also provides filtering and aggregation capabilities that can be used for preliminary co-registration of multi-stream data into probable „tracks" of attributes extracted from multiple sources that have a high probability of referring to the same entity or event. For example, if geo-encoded twitter messages mention a building fire at a given time and location and an acoustic sensor array determines that an emergency siren mounted to a vehicle is moving in the direction of that fire, then correlation in time, location, and semantic metadata (in this case, description of a fire and an emergency vehicle); then hard and soft data have effectively corroborated each other. With twitter data alone, it is difficult for a machine process (or even a human) to ascertain the veracity and relevance of an open source message. Similarly, hearing a siren alone doesn't provide information about the emergency vehicle's destination or the nature of the threat that it is responding to. However, extraction, correlation and fusion of these modalities can provide details that are impossible to obtain with either modality alone. Implementation details of our approach to this process are available in [26].

3.4. User Interactions

Careful consideration of user interaction is especially important in the case of a visually impaired user. The use of standard headphones for delivering sonified information to assist in navigation is impractical because headphones cover the ear canal and block critical ambient auditory cues. Alternatively, bone-conduction headphones, such as Georgia Institute of Technology's „bonephones" (see Figure 2) provide effective delivery of auditory displays [27] without blocking ambient noises.

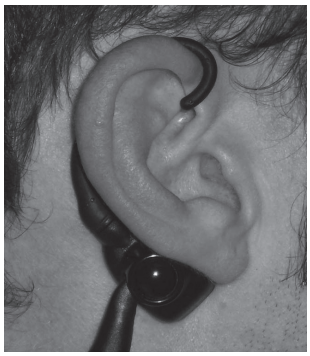

Figure 2: GA Institute of Technology's „bonephones" provide transmission of auditory displays without coving the ear canal from [28].

3.5. Test and Evaluation

The test and evaluation of hard and soft fusion processes is particularly challenging due to the subjective nature of human reporting, difficulty in determining the accuracy and veracity of open source data, and the relatively recent emergence of the hard and soft fusion paradigm. While techniques for calibrating and testing the performance of physical sensors are well established, human information reporting capability is subject to a large degree of variation based on a broad variety of factors (see [29] for examples). Although utilization of soft data sources adds tremendous potential for augmenting existing assistive technologies, it is critical to be mindful of the intrinsic uncertainties of this data and to design systems accordingly.

4. A SONIFICATION APPROACH FOR HARD/SOFT FUSION

New advances in integrating sonification with visualization and storification were also introduced by this research effort. The techniques proposed in [30] could be adapted for use by the visually impaired or temporarily blinded (by thick smoke, for example) in a manner that allows the user to dynamically modulate information presentation between sonification and visualization.

Additionally, we are experimenting with audible representation of fused hard and soft data that provides the user with brief, contextually modulated sounds that are informative of possible danger or other relevant notifications without being intrusive. The technique under development uses multi-agent software (MAS) techniques to add functionality that could augment the navigation beacon tone in SWAN (or similar tools) with another intermittent tone that would alert the user to relevant threats or alerts derived from soft data sources described above. For example, if the user were walking across town to a medical appointment, the software agents would monitor news, weather, police reports, and public posts on social media sites such as twitter. The user would only be alerted (via tonal variation) if the system deemed the alert likely to be relevant, timely, and geospatially close enough to affect the user. After receiving the subtle alert tone (indicating estimated urgency and alert type),

the user would have the option of either hearing the full-audio transcription or ignoring the alert.

Sonification is a natural compliment to hard and soft information fusion because it can enable the listener as a "human-in-the-loop" to recognize subtle relationships and correlations spanning multiple input streams and modalities. Although certain categories of hard/soft data can be fused automatically using techniques described above, there is another potential class of application in which physical sensors, distributed open source human observers (e.g. twitter users), artificial intelligence tools, and advanced HCI techniques all serve to deliver the proper elements of information to a human user at the ideal time and in the optimal format to enable them to use the exquisite information fusion capabilities of the human cognitive and sensory system to provide comprehensive real-time situational awareness despite physical, sensory, or cognitive impairments.

5. CONCLUSION

The interdisciplinary integration of hard and soft information fusion and sonification with existing accessibility techniques and applications has great potential to improve the safety and convenience with which blind or otherwise impaired individuals can navigate between locations, recognize objects in the world around them, and interact with other people. Many of the techniques described here are already being successfully utilized individually in domains as diverse as disaster management [31], financial trading [32], and disease informatics [33]. When properly integrated, these techniques, in conjunction with the exponentially increasing power of mobile devices and the rapid growth of freely available online information sources, create a "perfect storm" of converging factors to enable the next generation of accessibility tools.

6. REFERENCES

[1] D.J. Natale, R.L. Tutwiler, M.S. Baran and J.R. Durkin, "Using full motion 3D Flash LIDAR video for target detection, segmentation, and tracking", *Image Analysis & Interpretation (SSIAI), 2010 IEEE Southwest Symposium on*, 2010, p. 21-24.

[2] J. Wilson, B.N. Walker, J. Lindsay, C. Cambias and F. Dellaert, "Swan: System for wearable audio navigation", *Wearable Computers, 2007 11th IEEE International Symposium on*, 2007, p. 91-98.

[3] http://sonify.psych.gatech.edu/research/swan/

[4] S.M. Oh, S. Tariq, B.N. Walker and F. Dellaert, "Map-based priors for localization", *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3, 2004, p. 2179-2184.

[5] B.N. Walker and J. Lindsay, "Navigation performance with a virtual auditory display: Effects of beacon sound, capture radius, and practice", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 48, 2006, p. 265-278.

[6] B.N. Walker and J. Lindsay, "Using virtual environments to prototype auditory navigation displays", *Assistive Technology*, vol. 17, 2005, p. 72-81.

[7] J. Rimland, "Service Oriented Architecture for Human Centric Information Fusion", *Distributed Data Fusion for Network Operations*, CRC Press, 2012.

[8] M. Bujacz, P. Skulimowski and P. Strumillo, "Naviton: A Prototype Mobility Aid for Auditory Presentation of Three-Dimensional Scenes to the Visually Impaired", *Journal of the Audio Engineering Society*, vol. 60, 2012, p. 696-708.

[9] F. Zhigang and L. Ting, "Audification-based Electronic Travel Aid system", *Computer Design and Applications (ICCDA), 2010 International Conference on*, vol. 5, 2010, p. V5-137.

[10] W. Heuten, D. Wichmann and S. Boll, "Interactive 3D sonification for the exploration of city maps", *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, 2006, p. 155-164.

[11] P. Maes, "Agents that reduce work and information overload", *Communications of the ACM*, vol. 37, 1994, p. 30-40.

[12] P.J. Modi, M. Veloso, S.F. Smith and J. Oh, "Cmradar: A personal assistant agent for calendar management", *Agent-Oriented Information Systems II*, 2005, p. 169-181.

[13] K. Myers, P. Berry, J. Blythe, K. Conley, M. Gervasio, D.L. McGuinness, D. Morley, A. Pfeffer, M. Pollack and M. Tambe, "An intelligent personal assistant for task and time management", *AI Magazine*, vol. 28, 2007, p. 47.

[14] M.S. Uddin and T. Shioyama, "Detection of pedestrian crossing and measurement of crossing length-an image-based navigational aid for blind people", *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, 2005, p. 331-336.

[15] C. Estrada, J. Neira and J.D. Tardós, "Hierarchical SLAM: real-time accurate mapping of large environments", *Robotics, IEEE Transactions on*, vol. 21, 2005, p. 588-596.

[16] X. Jie, W. Xiaochi and F. Zhigang, "Research and Implementation of Blind Sidewalk Detection in Portable ETA System", *Information Technology and Applications (IFITA), 2010 International Forum on*, vol. 2, 2010, p. 431-434.

[17] A. de Campo, "Toward a data sonification design space map", *Proceedings of the International Conference on Auditory Display (ICAD)*, 2007, p. 342-347.

[18] J. Rimland, "Hybrid Human-Computing Distributed Sense-Making: Extending the SOA Paradigm for Dynamic Adjudication and Optimization of Human and Computer Roles",

A Dissertation in Information Sciences and Technology, The Pennsylvania State University, 2013.

[19] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*, O'Reilly Media, Incorporated, 2008.

[20] http://www.alchemyapi.com

[21] J. Rimland, J. Graham, G. Iyer, R. Agumamidi and S. Venkata, "JDL Level 0 and 1 algorithms for processing and fusion of hard sensor data", *SPIE Proceedings*, 2011.

[22] M. Botts, G. Percivall, C. Reed and J. Davidson, "OGC® sensor web enablement: Overview and high level architecture", *GeoSensor networks*, 2008, p. 175-190.

[23] M. Botts, G. Percivall, C. Reed and J. Davidson, "OGC sensor web enablement: Overview and high level architecture (ogc 07-165)", *Open Geospatial Consortium white paper*, vol. 28, 2007.

[24] A. Muys, "Building an Enterprise-Scale Database for RDF Data", *Netymon technical paper*, 2006.

[25] A. Schmeder, A. Freed and D. Wessel, "Best practices for open sound control", *Linux Audio Conference*, vol. 10, 2010.

[26] J. Rimland, M. McNeese and D. Hall, "Conserving Analyst Attention Units: Use of Multi-agent Software and CEP Methods to Assist Information Analysis", *Proceedings of SPIE 2013*, 2013.

[27] B.N. Walker and R. Stanley, "Thresholds of audibility for bone-conduction headsets", *International conference on auditory display, Limerick, Ireland*, 2005, p. 218-222.

[28] http://sonify.psych.gatech.edu/research/bonephones/index.html

[29] J. Rimland, D. Hall and J. Graham, "Human cognitive and perceptual factors in JDL level 4 hard / soft data fusion", *Proceedings of SPIE*, 2012.

[30] J. Rimland and M. Ballora, "Beyond visualization of Big Data: a multi-stage data exploration approach using visualization, sonification, and storification", *Proceedings of SPIE*, 2013.

[31] N. Bessis, E. Asimakopoulou and F. Xhafa, "A next generation emerging technologies roadmap for enabling collective computational intelligence in disaster management", *International Journal of Space-Based and Situated Computing*, vol. 1, 2011, p. 76-85.

[32] G. Cugola and A. Margara, "Processing flows of information: From data stream to complex event processing", *ACM Computing Surveys*, 2011.

[33] M. Salathé and S. Khandelwal, "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control", *PLoS computational biology*, vol. 7, 2011, p. e1002199.