Supporting audiography: Design of a system for sentimental sound recording, classification and playback

Tijs Duel, David M. Frohlich, Christian Kroos, Yong Xu,
Philip J. B. Jackson & Mark D. Plumbley
University of Surrey, Guildford GU2 7XH, UK
t.duel@surrey.ac.uk

Abstract. It is now commonplace to capture and share images in photography as triggers for memory. In this paper we explore the possibility of using sound in the same sort of way, in a practice we call *audiography*. We report an initial design activity to create a system called *Audio Memories* comprising a ten second sound recorder, an intelligent archive for auto-classifying sound clips, and a multi-layered sound player for the social sharing of audio souvenirs around a table. The recorder and player components are essentially user experience probes that provide tangible interfaces for capturing and interacting with audio memory cues. We discuss our design decisions and process in creating these tools that harmonize user interaction and machine listening to evoke rich memories and conversations in an exploratory and open-ended way.

Keywords: Audiography, sound souvenirs, design, machine listening

1 Introduction

We have all become familiar with capturing and sharing photographs as triggers for memory, reminiscing and storytelling. In fact, modern domestic photography is as much about communication as it is about imaging and memory [12]. The popularity of photography is evident if we look at the volume of photographs uploaded to social media sites every day. Facebook for example saw over 350 million pictures uploaded daily in 2013 [1]. However, this volume of photographs may be considered a problem for remembering as it becomes hard to retrieve photos relating to particular events, and easy to forget the photographs themselves [13, 7]. Furthermore photographs only relate to our visual memory for events.

In this paper we consider an alternative medium for the triggering and sharing of memories based on hearing rather than seeing, sound rather than image, and face-to-face sharing rather than online posting of media. We call this activity '*audiography'* to contrast it with photography and videography, and connect it with insights from our own prior work on audiophotography [6] and that of others exploring the sentimental properties of sound for memory and communication [2, 4, 10].

These studies show that audio recordings provide unique experiences and qualities for remembering compared to their visual counterparts, but that audio is more difficult to index and navigate in a non-visual way. For example, Bijsterveld & van Dijck [2] point out that professional music has always had nostalgic value for owners along with

a variety of often accidental sound recordings made in the cassette tape era. Their collection of recordings show how these 'sound souvenirs' are captured and shared in different contexts. The same discovery was made in the first study of the domestic soundscape by Oleksik et al [9] when asking families to record various sounds of family life. The families enjoyed this activity and asked to keep certain recordings which the authors termed 'sonic gems' because of their function in triggering precious memories. Dib et al [4] supported this activity directly in a field trial of digital Dictaphones given to families for the purpose of remembering. This uncovered a host of values for sentimental sound recording, including the immersive nature of sound for transporting people back in time, its ambiguity in making space for imagination in recall and discussion in conversation, and its spontaneity and uncontrollability compared to images which can be more easily 'posed' than recordings.

These latter studies acknowledge the difficulty of organising and browsing digital sound recordings and suggest novel ways of making them tangible. Hence, Petrelli et al [11] describe the design and use of a device called FM Radio (short for Family Memory Radio) through which families could 'tune in' to sound recordings clustered by time, type or favourites. Oleksik & Brown [10] describe the design and use of a system called Sonic Gems for recording and storing sentimental sound recordings in the form of pebble-like capsules in a bowl. Each device showed promise in supporting a new practice of sentimental sound recording and playback, but neither was designed to scale. Families had to classify recordings manually for playback through FM Radio and the unclassified Sonic gems would soon fill the bowl and become unmanageable. Linking sound recordings to photographs through 'audiophotography' provides a more sustainable solution to managing audio collections visually [5, 6]. However, the primacy of visual stimulation in an audiophoto may overpower some of the psychological properties of audio for remembering. So a requirement exists for a more manageable and scalable way of classifying and curating sentimental sound recordings.

In the rest of this paper we describe a design exploration to do this, drawing on machine hearing algorithms being developed on the *Making Sense of Sounds* project. This is a large multidisciplinary project at the Universities of Surrey and Salford in the UK, attempting to recognise and classify scenes from their sonic properties. Engineering work at Surrey and psychoacoustic work at Salford are being combined to deliver a method of automatically tagging naturalistic sound recordings with psychologically meaningful labels in a taxonomy of sound types. We believe this could be a critical enabling technology for sustainably managing sentimental sound collections and making them available for powerful new experiences of remembering and discussing the past.

2 Approach

To explore this design space we used a research-through-design approach to design a working prototype system called *Audio Memories*. This comprises separate audio recorder and playback devices operating as a pair. Rather than using standard consumer electronic platforms such as a smartphone and computer, custom appliances were

created. This was done to rethink the design of such devices for this application and with sound as the dominant medium.

For example, we attempted to make the interfaces to the devices both tangible and screenless, excluding the conventional method of searching for media by text label input. This forced us to develop new methods of query using physical buttons and control knobs, with some level of unpredictability and serendipity of selection. We also wanted to make the devices attractive and accessible to young families with children as the target market, since they were shown in some of the previous studies above to be interested in sound souvenirs.

This paper is written as a design case study, culminating in the description of a working prototype. It essentially describes our design iterations over time for the recorder and player devices separately. In a future paper we will describe a full trial of the system with target families, to evaluate the effectiveness of the prototypes in meeting the design requirement for a scalable practice of audiography

3 Recorder evolution

Fig. 1. Images of the main iterations of the recorder.

2.1 Recorder I

Dibs' and Oleksik's, research participants requested tools to create short yet significant clips from their longer recordings, confirming the desire for an event-focused reminiscing. To answer this need we fixed the length of each recording to 10 seconds exactly. In this way the user would be encouraged to treat a recording at the onset actively as an (physically ephemeral) event instead of an indistinct variable-length documentation, but the disruption of any ongoing personal or environmental interaction would be minimized. Fixing the length of the clip also rendered pause and stop button superfluous; a single button would accomplish the entire process and it would not even be necessary to pay visual attention to it. A working mono-recorder was built as proof of concept (see Figure 1.i).

Besides investigating interaction and hardware we also started exploring the medium of ten second clips. We asked people linked to our group to start recording such clips and this resulted in a small library of audio recordings. We found that 10 seconds was largely sufficient to capture meaningful sound mementos.

2.2 Recorder II

The second design iteration introduced new core functionalities and explored design details of the casing (see Figure 1 ii & iii). The core hardware was upgraded to enable 16 bit 44.1 kHz stereo audio. The recordings were meant to capture the sonic experience of the original situation as closely as possible and using high-quality audio was deemed indispensable despite the larger processing and energy consumption footprint.

To give the user some instant control over the recorded content without compromising the overall idea of a minimal interfering device, a smaller ‘delete’ button was added next to the 'record' button. By triple-tapping this button the user is able to permanently remove the most recent recording.

We also took into consideration adding audio playback capabilities to the recorder, but after careful deliberation we decided against it. Firstly, the user’s engagement in the ongoing situation should not be interrupted by the temptation to play back the recorded sound instantly. Secondly, the anticipation of listening to the recordings later on the playback device was considered to be a desirable quality on its own.

Finally, GPS localisation was implemented to add location meta-data to the recordings including also a time stamp. These meta-data provide initial means to index and navigate the accumulating audio collection.

2.3 Final recorder prototype

The final prototype for the recorder is shown in Figure 3 below. The electronics were complimented with a vibration motor providing haptic feedback as to when a recording is started and completed. Furthermore, we improved the battery charger to reduce downtime. Most improvements were made on the case. The addition of a bright orange gave the design a friendlier and more playful nature. The sides are made using 3D printing improving its structural strength. The top and bottom required more accurate machining and are made through lasercutting Perspex. Using these rapid prototyping techniques, we improve reproducibility of the design. The intimate nature of audio recording raised issues of confidentiality. For the most secure data transfer to the player device we introduced a removable micro SD card at the bottom of the device.

3 Player evolution

4.1 Early conceptual designs

Fig. 2. Some examples of made concepts for the player. Each concept has its own way to navigate audio collections.

To address the design challenges of digital archiving we adopted a more open-ended design approach; exploring multiple concepts through rapid prototyping. The resulted exploration can be classified in three categories: The first category is **GPS oriented searching** (see Figure 2.i); moving the speaker cones user could dowse for audio recordings. Finding sounds through this method could proof to be hard as the locations of recordings are not distributed evenly through space.

The second type of concept focusses on **chronological oriented searching** (see Figure 2.ii). The concept featured a vertical pole representing the stack of recordings in chronological order (newest on top). The speaker visualizes it relative position in time. This concept of browsing the collection definitely holds its merits however, the prototypes was considered awkward to operate as it moved every ten seconds.

The final group of concepts explored the idea of **co-operative reminiscing** (see Figure 2.iii). Multisided interfaces enabled multiple users to operate the device simultaneously triggering a multitude of audio memories to create layered soundscapes. A technical demonstrator was realized using MAX/MSP and a surround sound system. The layering provided interesting reminiscing experiences, through introducing serendipity. One problem discovered in the audio experiments here was that some sound clips 'clashed' with each other. Typically sounds of the same type such as music or voiceover were difficult to listen to simultaneously. In addition, the short nature of the audio memories made it hard to effectively browse and layer sounds. Playback would move onto the next clip in the stack while users were deciding whether that was the clip they wanted to listen to.

3.1 Final Player prototype

The final prototype blended features and functionalities of its prior explorations into a single design. It provides a tangible interface to browse one's digital audio mementos with up to two clips playing simultaneously (see Figure 3).

Fig. 3. The final designed system

The bottom left of the device features four tag buttons. These buttons can be used to select the content of the clips; mechanical sounds, nature sounds, speech and music. Combinations can be made as well. The categories are based on the audio taxonomy of everyday sounds by Bones et al [3]. When the SD card of the recorder is inserted into the player it automatically downloads, analyses and labels the recordings. An audio classification method will label each clip with one (or more) of the four tags based on its dominant content.

At the center of the interface is the Haptic Engine (H.E.). The H.E. is a dial that doubles as a loci of both input and output. While a clip plays the H.E. - similar to a record player – will slowly rotate, making a full revolution every ten seconds. Equipped with capacitive touch sensors it will disengage the motor and pause playback when held. Manual rotation enables users to browse their collection in time. Continuous rotation will make one jump through the collection in larger intervals. Users can keep track of their position within their collection through a 16x8 LED matrix depicting the date (in a MM-YY format). The LED matrix is situated behind the speaker cloth. When the LED's are off the matrix practically disappears.

At the back of the device is a rotary switch that enables the player to switch into layer mode. In this mode the player will analyse the track that is playing based on user input and layer it with a matching track played through a second speaker. This emulates a simplified spatial audio effect.

4 Discussion

Although we cannot present a tested solution for audiography, we have reported a design exploration culminating in a potential solution ready to test. This adopts a design approach based on tangible interaction with separate recording and playback devices that utilise time and GPS meta data inherited from the recorder. Further metadata in the form of sematic tags of sound type are added automatically by deep learning algorithms on the player. These allow searching for sounds by top level sound categories such as

mechanical, nature, speech and music, together with chronological position in a ‘clip stack’ that has emerged as an important concept approximating time.

Various design issues surfaced in exploring options for the recorder and player which may be common to other audio recording and archiving domains. Regarding the recorder we found that the removal of a screen and speaker on a dedicated device leads to a simple one button interaction for recording. The lack of ability to review sounds on the recorder creates anticipation of the playback experience on the second device, and encourages batch transfer and playback in specific episodes. Ethical issues about the capture and control of sounds led us to incorporate LEDs signaling recording activity on the recorder, and removable storage which is physically moved to the player and cannot be accessed wirelessly.

Regarding the player, we quickly found that it was easier to use time and sound type metadata in comparison to GPS location. Without a screen it was not possible to use a map interface for sound clips, and alternative methods of pointing the device itself in various directions were not intuitively obvious. Therefore a static tabletop device for shared use was chosen for development, using a small number of buttons and dials for tangible control. The possibility of mixing or layering multiple sounds at playback was explored experientially with 10 second clips recorded on a smartphone to test the approach. This revealed the difficulty of listening to too many sounds at once and encouraged the design of a secondary layer as an option to a primary one. Many design discussions for the final player centred around the user experience of searching for specific sounds versus exploratory listening to sound sequences and combinations from approximate time points. Here we were inspired by studies showing the value of serendipitous interaction with media, to allow the system to play sound continuously with approximate user ‘steering’ [7, 8].

In general, we found ourselves moving between design of different types and levels in order to resolve the emerging design issues. Conceptual design was necessary at the top level to define a design trajectory for each device, which then led us into explorations of form, functionality and media. Because both devices were for handling sound, we found it necessary to record sound samples at various points and experiment with the experience of playback in different sequences and combinations.

A final dynamic for our player interaction, was the division of labour between human and machine intelligence and action. We found ourselves in unfamiliar territory of combining artificial intelligence and tangible interaction, through the sound type (tag) buttons which could retrieve a sub-set of sounds by type through a physical action. The complexity of manually selecting a secondary sound clip to play with a primary one also led us to leave this to the machine, based on a simple algorithm for ruling out ‘clashes’. Whether users understand or enjoy the logic and experience of this kind of sound selection and playback remains to be seen in a user trial we plan to run next. For now we can say that machine listening and human listening can indeed be combined in a single device, to curate sounds from an archive in a scalable and interesting way.

Acknowledgements

This work was funded by a grant from the Engineering and Physical Sciences Research Council, as part of the Making Sense of Sounds Project (EP/N014111/1). We would like to thank our other colleagues on that project for their input to the design ideas in this paper.

References

1. Aslam S. (2018). Facebook by the numbers: Stats, demographics and fun facts. Omnicore online magazine. https://www.omnicoreagency.com/facebook-statistics/
2. Bijsterveld, K., & van Dijck, J. (Eds.). (2009). *Sound souvenirs: audio technologies, memory and cultural practices* (Vol. 2). Amsterdam University Press.
3. Bones, O. C., Cox, T. J., & Davies, W. J. (2016). Toward an evidence-based taxonomy of everyday sounds. *The Journal of the Acoustical Society of America*, *140*(4), 3266-3266.
4. Dib, L., Petrelli, D., & Whittaker, S. (2010). Sonic souvenirs: exploring the paradoxes of recorded sound for family remembering. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 391-400). ACM.
5. Frohlich, D. M. (2004). *Audiophotography: Bringing photos to life with sounds.* Springer Science & Business Media.
6. Frohlich, D. M. (2015). *Fast Design, Slow Innovation*: *Audiophotography ten years on.* Springer International Publishing.
7. Frohlich, D. M., Wall, S., & Kiddle, G. (2013). Rediscovery of forgotten images in domestic photo collections. *Personal and ubiquitous computing*, *17*(4), 729-740.
8. Odom, W. T., Sellen, A. J., Banks, R., Kirk, D. S., Regan, T., Selby, M., & Zimmerman, J. (2014, April). Designing for slowness, anticipation and re-visitation: a long term field study of the photobox. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1961-1970). ACM.
9. Oleksik, G., Frohlich, D., Brown, L. M., & Sellen, A. (2008). Sonic interventions: understanding and extending the domestic soundscape. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 1419-1428). ACM.
10. Oleksik, G., & Brown, L. M. (2008). Sonic gems: exploring the potential of audio recording as a form of sentimental memory capture. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1* (pp. 163- 172). British Computer Society.
11. Petrelli, D., Villar, N., Kalnikaite, V., Dib, L., & Whittaker, S. (2010, April). FM radio: family interplay with sonic mementos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2371-2380). ACM.
12. Sarvas, R., & Frohlich, D. M. (2011). *From snapshots to social media-the changing picture of domestic photography*. Springer Science & Business Media.
13. Whittaker, S., Bergman, O., & Clough, P. (2010). Easy on that trigger dad: a study of long term family photo retrieval. *Personal and Ubiquitous Computing*, *14*(1), 31-43.