Summer 2014

Personalized audio warning alerts in medicine

Todd Alan Papke
University of Iowa

PERSONALIZED AUDIO WARNING ALERTS IN MEDICINE

by

Todd Alan Papke

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Informatics (Health Informatics)
in the Graduate College of
The University of Iowa

August 2014

Thesis Supervisor: Professor Prakash Nadkarni

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Todd Alan Papke

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Informatics (Health Informatics) at the August 2014 graduation.

Thesis Committee: _____
Prakash Nadkarni, Thesis Supervisor

_____
Warren Boe

_____
Joann Eland

_____
Juan Pablo Hourcade

_____
Priyadarshini Pennathur

To Patty and Annabelle.

It matters not much now, but it might have significance in the long run.
Onward!

John Bennett

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor and mentor, Dr. Prakash Nadkarni. Not only did he insist that my research was doable, he believed that it needed to be done – under his mentorship I was able to turn an ember of an idea into a roaring good adventure. He is accessible, inspires confidence, and, above all, is compassionate – qualities I will hold dear as I move forward with students of my own. To my committee members – Drs. Hourcade, Eland, Boe, and Pennathur – your support, ideas, and commitment were invaluable in making this dissertation possible. Your viewpoints opened doorways and gave me insights I otherwise would have missed, and my work greatly benefited as a result.

To all my fellow graduate students – my friends – I treasured each and every moment you chose to share your world with me. Charisse, Si-Chi, Ray – it was an honor to watch you tackle your topics and metamorphosize into true researchers. I get excited knowing that you all are just getting started, and I smile when I think about your future having witnessed where you've been.

Finally to my wife and family, your love and support meant the world to me. Patty, your confidence in me never wavered, even when my confidence in myself ran aground. This dissertation would have forever been a dream without your love and support. You made the impossible, possible and the unreal, reality. You and Annabelle are my pride and joy.

ABSTRACT

Modern Electronic Health Record (EHR) systems are now integral to healthcare. Having evolved from hospital billing and laboratory systems in the 80's, EHR systems have grown considerably as we learn to represent more and more aspects of patient encounter, diagnosis and treatment digitally. EHR user interfaces, however, lag considerably behind their consumer-electronics counterparts in usability, most notably with respect to customizability. This limitation is especially evident in the implementation of audible alerts that are coupled to sensors or timing devices in intensive-care settings. The most current standard, (ISO/IEC 60601-1-8) has been designed for alerts that are intended to signal situations of varying priorities: however, it is not universally implemented, and has been criticized for the difficulty that healthcare providers have in discriminating between individual alarms, and for the failure to incorporate prior research with respect to "sense of urgency" as it applies to alarm efficacy. In the present work, however, we consider that there are more effective means to allow a user to identify an alarm correctly than "sense of urgency" response.

This thesis focuses on the problem of correct identification of alerts: what happens when a human subject is allowed to create or designate (i.e., personalize) one's own alerts? Given the ubiquity, low costs and commoditization of consumer-electronics devices, we believe that it is just a matter of time before such devices become the norm in critical care and replace existing, special-purpose devices for information delivery at the point of patient care.

We built a tool, PASA (Personalized Alerts Study Application), that would allow us to capture and edit sounds and orchestrate studies that would contrast any two types of sounds. PASA facilitated a study where study participant's responses to "personalized" sounds were contrasted with sounds that meet the ISO/IEC 60601-1-8:2012 standard.

We performed two sub-studies that contrasted responses to two banks of 6-alerts and 10-alerts. The 6-alert study was repeated with the same subjects after two weeks without training to measure recall. We observed that accuracy, reaction time, and retention were significantly improved with the personalized sounds. For example, the median errors for the 6-alert baseline study were 4 for personalized vs. 27 for standard alerts. For the 6-alert repeat study it was 7 vs. 43. The median for the 10-alert study was 1 for personalized vs. 55 for standard alerts. Accuracy for recognition, while remaining constant for personalized alerts, degraded considerably for standard alerts as the number of alerts increased from 6 to 10.

We conclude that personalization of alerts may improve information delivery and reduce cognitive overload on the health care provider. This potential positive effect at the point of patient care merits further studies in a clinical or simulated clinical setting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The field of healthcare poses a vast and exciting variety of computational challenges. Electronic health record (EHR) systems have the potential to evolve into becoming the hub for all aspects of health care delivery. Going beyond patient data, gene sequences, and imaging information, EHR systems of the future may provide real-time evidence-based decision support that will vastly improve patient safety, by providing "best practice" information needed for treatment of disease.

The most important component of the "system" is the healthcare provider, who is required to perform specific actions on the patient after mentally processing diverse inputs. An important category of inputs, especially in critical-care settings, are audio alerts, which are the focus of the present proposal. Personalizing the alerts to the provider's preferences and previous level of skill and training can enable a more effective level of performance, as measured by faster response times and more accurate responses, especially during critical situations and when multiple alerts occur simultaneously. The potential positive impact on patient safety therefore requires that new ways of personalizing audio alert delivery be studied, through a quantitative experimental design that contrasts personalized alert delivery with the traditional "one size fits all" approach.

This thesis will explore audible alerts in the context of medicine. We will look at previous research on audible alerts, and draw lessons from it. We will provide a fresh experiment design, one that uniquely builds upon elements from the research corpus that are applicable to personalized alerts and provides data that may give some insight into whether personalized alerts have the potential to improve patient safety.

This chapter will discuss the significance of this research, hypotheses that we would like to test, why we should consider research that explores incorporating

personalized alerts in to the healthcare system, and limitations of our research methodology.

Chapter 2 will look at previous research on audible alerts within medicine; methods, topics of focus, and limitations. Chapter 3 will focus on our research on personalized alerts with an overview of methodology and tools, and Chapter 4 will present our results.

1.1 Significance of Study

This research will advance the science of audible alert design by establishing a study on recognition, response and retention of sounds created by a group of subjects who have been enabled to create their own alert sounds. This approach should result in a unique contribution by examining a new modality for the design and delivery of audio alerts. This would serve as a basis for research that would make personal audio alerts a reality in clinical settings.

1.2 Hypotheses

Hypothesis #1: Users will respond to personalized alert sounds more accurately than to "standard" medical alert sounds.

Hypothesis #2: Users will respond to personalized alert sounds more rapidly than to "standard" medical alert sounds.

Hypothesis #3: Retention of personalized alert sounds will be superior to "standard" medical alert sounds. In other words, people who are allowed to create and train on alerts with their own alert sounds will remember them better than if they trained on alerts with the standard sounds currently in use.

1.3 The Practical Importance of Personalizing Alert Sounds in Healthcare

In the field of healthcare, the area of intensive patient care is especially challenging because taking care of seriously ill patients requires specialized healthcare provider skills at both the physician and nursing levels. Intensive-care-unit (ICU) settings

have a high ratio of providers to patients, because patients' physical status may change from stable to critical within minutes, and, consequently, they are required to be monitored around the clock. To facilitate such monitoring, patients are connected to a variety of biological sensors that monitor various essential functions, such as heart rhythm, blood pressure, oxygen saturation and respiratory rate. Other sensors detect potential problems with mechanical devices that are used to administer therapeutic measures such as respirators, intravenous fluid pumps and venous and arterial catheters. Yet other devices are timer-based, prompting the nurse, for example, to change the patient's position periodically to prevent pressure ulcers in immobilized or unconscious patients. Finally, the patient may also summon the provider to their bedside by pressing a call button.

Some sensors (e.g., physiological parameter monitors) provide feedback through visual displays. Given, however, that visual displays require the provider to be looking at them continually in order to serve their purpose, all sensing devices must also provide auditory feedback (an "alerting sound"), which is triggered by an abnormal physiological pattern such as a ventricular premature beat, an out-of-range value such as a very rapid or very slow heart rate, or a timer. Given the multitude of sensors, it is important for the provider to be able to discriminate immediately and accurately between the various sources of sounds. One would expect, therefore, that each sensor's sounds would be configured to be sufficiently distinctive to allow ready differentiation.

In practice, this is not necessarily the case. At University of Iowa Hospitals and Clinics (UIHC), for example, all devices produce alerting sounds with a fixed frequency of 4000 Hz. Individual device alerts are distinguished by variations in the duration of a beep, the duration of silence between beeps, or by varying sequences of beeps (e.g., short beeps vs. long beeps, in the fashion of Morse code). Informal interviews with the ICU nurses indicate that it takes a minimum of three to four weeks to recognize alerts reliably. A practical issue is that the nurses do not undergo one or more training sessions that

include the full range of alerting sounds that they may be expected to encounter: learning is achieved by encountering an alert in an actual patient-care setting and asking colleagues what it indicates. As a result, even during the four-week phase, there may be an alert (or alerts) that a given nurse may never encounter because the particular device is not triggered by an abnormal patient parameter.

Despite these issues, the system works currently, largely because of the large nurse-to-patient ratio under normal circumstances of patient load, so that there is plenty of skilled-personnel backup, and the number of alerts during a given time period per nurse stays relatively manageable. One can very well imagine that the system would work less well if patient load were suddenly increased – e.g., because of a large-scale emergency or disaster situation. Here, the shortage of skilled staff might require recruitment of individuals – e.g., medical or nursing students, or even volunteers and relatives of patients – who need to learn to discriminate between a set of not particularly distinctive or memorable sounds, with very little time to accomplish learning.

Toward this goal, a significant body of research has been devoted to the issue of devising "standard" alert sounds in ICU and anesthesiology practice that have a universal meaning across medical contexts – e.g., high vs. medium vs. low priority – and that are also sufficiently memorable to be learned quickly and retained and responded to with high accuracy.

Due to the rapidly increasing power, ease of use, lowered costs and consequent ubiquity of electronic technology and connected devices, however, much of this research may well have become obsolete, as discussed in the next chapter. Users of a technology differ greatly in cultural and professional background, interests, and skills. Consequently, a search for the perfect universal set of "standard" sounds may well be quixotic. For almost a decade, vendors in fields as diverse as computing hardware and software, consumer electronics (e.g., the iPhone and Android phones), e-Commerce (e.g., Amazon), social media (Facebook), Web-based media delivery (Netflix) and even

holiday resorts (e.g., Disneyland/ Disney World) have realized the importance of product customizability, to enable personalization of the user experience.

With respect to audio, our creation, consumption and organization of sounds, as well as responses to sounds, have changed significantly since the time much of the research on audio alerts was performed. When using their personal electronic devices, end-users combine sound clips, song fragments, or voices in unique ways that provide alerts with highly contextualized meaning, such as a specific ringtone for a specific relative, or a song that plays when we have run our tenth mile. This technology is starting to appear in more intimate objects, such as devices that will seem to preemptively sense location and intent, as they provide sensory augmentation with information that is contextualized situationally. Google Glass is an interesting early example.

Healthcare technology, which at present continues to lag behind consumer technology considerably with respect to cost, usability, and commoditization, can benefit similarly from personalization. As technology and culture align, alert systems that will be ubiquitous and highly personalized with respect to situation, location, and role of the healthcare practitioner will be possible. It is important to explore research design for personalized alerts so that a research corpus can be formed that will more effectively influence future standards for various aspects of healthcare-provider-computer interaction.

1.4 Scope and Limitations of the Present Work

This research is limited to human response to auditory alerts, with a focus on personalized alerts. The human subjects tested all exhibited common abilities and limitations, and the only specific ability that we were interested in identifying was whether or not the subject had formal music training, as that has shown to correlate to better tone recognition in previous research [1]–[3]. Our research does not utilize all types of audible alerts, only a subset of those that adhere to the current international standard, ISO 60601-1-8. Further, while this study was performed in the context of health

care, it only serves as a model by which personalized alerts can be contrasted with other types of alerts. Much like previous research in audible alerts, an actual clinical trial in an intensive-care unit is beyond the current addressable scope of this work. However, the tools built to enable the work described in this thesis have been designed so that future studies that utilize clinical simulations may be considered without major re-engineering of the toolset.

This thesis is not intended to address the problems of alarm accuracy, or alarm/alert fatigue. All devices that generate alarms have a certain proportion of false positives (an alarm is generated when not necessary) and false negatives (failure to generate an alarm when necessary): alert fatigue is due to excessive false positive alerts. This problem is not limited to auditory alerts – excessive alerts, especially for drug-drug interactions, are a common complaint for users of Computerized Physician Order Entry (CPOE) systems.

The multiple root causes of excessive false-positive alerts have been explored extensively in the informatics literature, such as acting on incomplete information about the patient, excessively simplistic algorithms that consider only a single input rather than integrating multiple inputs, and not allowing for physiologic variations. There is scope for personalization here: the level of alerting/prompting that might be acceptable to trainees in their first few weeks or months of work would be unacceptable to an expert and highly experienced healthcare provider. Such customization, however, involves algorithmic, knowledge-representation and database-content issues that are beyond the scope of the present work.

CHAPTER 2
PREVIOUS WORK

This chapter provides background into previous work on audible alerts. We begin Section 2.1 with a description of the search process through which the research corpus was obtained. Section 2.2 is a literature review where we start with some information on alarm fatalities and specific instances of death that has resulted in problems with existing alarms in section 2.2.1 In Section 2.2.2 we will examine the reasoning, methodology, and findings of previous research in the field. Section 2.3 provides a summary and discussion of observations and limitations of past work.

2.1 Methodology for Corpus Acquisition

Once the topic of research was identified, a systematic review was carried out of prior research on audible alerts. Google Scholar, PubMed, ACM Digital Library, and Questia served as starting points. Articles of interest were identified and obtained through any electronic database providing access to abstracts and full articles - in many instances a number of research portals had to be explored before a comprehensive version of a specific publication could be found. Searches were done systematically beginning with specific key words, e.g., “audible alerts”, “alarms”, “sense of urgency response”, “alarm fatigue”, “alarm sounds”, “auditory warning”, “clinical alarms”, “alarms and medical equipment”, “psychoacoustics”, “cognition of sound”, “earcons”, “auditory icons”, “ambient noise and alarms”, “alarm retention”, “perceived urgency”, “false alarms”, “sound attributes”. Citation links were examined and followed rigorously, especially where a citation appeared more than once across the journal collection.

When electronic versions of articles were not available, an attempt to locate the associated journals was first made at The University of Iowa libraries, and, if necessary, procured through interlibrary loans. Articles were filtered with respect to applicability to audible alerts and their application in medicine. Also, with respect to early audible alert

research, an original typed manuscript was obtained in an effort to clarify assessments made in research articles critiquing early work that were somewhat contradictory in nature.

For a historical perspective, articles that were most commonly cited were reviewed repeatedly in an effort to identify and isolate the prevailing research design; the reasoning, methodology, and results from each set of experiment(s) performed in advancing the field of audible alerts.

Of an original corpus in excess of 1200 articles, 261 were selected based upon relevancy to audible alerts and their application in medicine. Out of the 261, there was a high degree of similarity. Therefore, a smaller selection was utilized in an effort to focus on the more relevant topics for this thesis.

2.2 Literature review

In the early 80s the medical community began questioning the extant design of audio alerts. Kerr et al. looked at the growing number of alarms in the intensive care unit, and found that "the provision of audible alarms is eminently logical when considered from viewpoint of the individual item of equipment but problems arise when several such items are collected together around a patient" and that "six or more sounds emanate from each patient" [4]. At that time most audible signals consisted of continuous or rapid intermittent tones of constant pitch whose loudness was often uncontrollable, to ensure attention across a moderately noisy room. Most equipment needed to be placed close to the patient's bed; therefore exerting "their greatest effect on the patient" [4].

Even experienced intensive care nurses, capable of distinguishing signals emitted by different types of equipment, admitted to difficulties when several alarms with similar sounds were generated simultaneously. Less experienced nurses became "confused and unable to react rapidly when several alarms sound simultaneously or in quick succession" [4]. The stress in an ICU, combined with the fact that the number of alerts occurring simultaneously is a function of the number of patients in a ward, caused the nurses to

eagerly silence audible signals. As a result, the alarm designers were forced to go to considerable lengths to circumvent these tendencies. In 2010, Siebig et al., [5] examined the clinical and technical validity of ICU alarms to determine the potential value of new algorithms to maintain high sensitivity while improving alarm specificity. They noted that the correct identification of critical alarms by medical staff has been historically reported as low as 40% [6].

2.2.1 Problems with Alarms

Inability to recognize alarms, often resulting from alarm fatigue has caused deaths. The Food and Drug administration received 862 death reports associated with alarms in the seven years that ended in 2012 [7]. Reporters for the Boston Globe, through public records requests, found at least 11 deaths in Massachusetts since 2005 linked specifically to lack of response, or inadequate response to alarms on cardiac monitors in hospitals [8].

The Joint Commission, which accredits hospitals, received 98 reports of alarm-related incidents of which 80 resulted in deaths in the 42-month period ending June 2012. “Those voluntary reports are a gross undercount, says the commission, which estimates that there were close to 1,000 alarm incidents in which patients died or were injured, or faced those risks.” [7],[9].

In 2013 The Joint Commission directed facilities to make alarm safety a top priority or risk losing their accreditation. The commission is requiring hospitals, starting in January, to identify the alarms that pose the biggest safety risks by unnecessarily adding noise or being ignored. By 2016, hospitals must decide who has the authority to turn off alarms [7].

At John Hopkins, an alarm task force found that the average number of alarms that sounded per bed per day in one ICU was 771 [7].

The ECRI Institute, a Pennsylvania-based patient-safety organization, listed alarm hazards as the No. 1 issue on its annual list of the top 10 health-technology dangers for 2012 and 2013 [7],[10].

Some specific examples:

In 2009, At St. Elizabeth's Medical Center in Brighton, MA, a patient's monitor cable became loose and the machine repeatedly sounded a low-pitched beep. Nurses said they did not hear the alarm. They did not discover the patient had stopped breathing until it was too late [8].

In August, 2010, a 60-year-old man died at UMass Memorial, Massachusetts, after alarms signaling a fast heart rate and potential breathing problems went unanswered for nearly an hour [8].

In January, 2009, Linda Knyff, 47, died at Massachusetts General Hospital in Boston after the tube that had been surgically inserted into her trachea to help her breathe became dislodged. Because of an apparent malfunction, her cardiac monitor was not setting off alarms at the central monitor at the nurses' station. Nurses would have been alerted to the malfunction by icons and alarms on the central monitor and on hallway signs [8].

In January, 2010, 89-year-old man died at Mass. General. Ten nurses on duty could not recall hearing the beeps at the nurses' station or seeing scrolling ticker-style messages on three hallway signs that would have warned them as his heart slowed and finally stopped over a 20-minute span [8].

In Maryland an elderly woman's heart-rate monitor beeped unnoticed by staff for several minutes. Her nurse was preparing medication in another room where she could not hear or see the patient or the monitoring system. (The hospital had cut the position of the technician whose job was to monitor alarms at a central station.) By the time the nurse responded, the patient had suffered irreversible damage, and she died shortly afterward [7].

The parents of Mariah Edwards won a $6 million malpractice settlement after their 17-year-old daughter died in 2013 following a tonsillectomy at a Pennsylvania surgery center. After the surgery, the high school junior was given a potent painkiller that slowed her breathing. By the time nurses checked on her 25 minutes later, she had suffered profound and irreversible brain injury. She died 15 days later. A nurse said in her deposition that the alarm on the respiratory monitor was muted, said Joel Feller, an attorney for the family. After Edwards's death, the center announced several changes, including that alarms would no longer be muted [7].

2.2.2 Audible Alarm Research

2.2.2.1 Sense of Urgency Response

Insights into sense of urgency response, and experiment design that might identify culturally agnostic aggregations of sound attributes began when Patterson published guidelines for auditory warning systems on civil aircraft [11]. In 1985 his team from The Applied Psychology Unit from Cambridge worked in conjunction with the Institute of Sound and Vibration research from Southampton to design alarm sounds for British nuclear and military applications. He made recommendations for auditory warning systems for workload environments [12], and then published a seminal work on medical alerts, which provided guidelines that explained the spectral characteristics required to make a pulsive sound that is distinctive and resistant to masking unexpected noise sources[13].

Patterson's sound categories formed the basis of an ISO committee draft pertaining to alert sounds, ISO/TC 1/SC3, and remained influential nine years later when ISO ISO:9703-2:1994, "Anaesthesia and respiratory care alarm signals – Part 2:Auditory Alarm signals" was published [14]. These categories also were used for sounds specified by IEC/ISO 60601-1-8:2006, the most current international standard for audible alerts in medical devices.

Category-specific sounds were constructed for general, oxygenation, ventilation, cardiovascular, artificial perfusion, drug administration and temperature. Each of these unique sounds was provided in Caution and Emergency forms, to be repeated at the same intervals as the general sound (30s and 15s respectively).

All sounds were represented with two levels of urgency encoding with a third encoding representing general information sounds. Each alarm burst would have a period of silent time so that verbal staff communication could take place without interruption. The sounds would utilize highly distinctive, tonally complex sounds that would not easily be confused with anything else, and guidelines were provided for designing sound pulses that would provide this tonal complexity that specified pitch, harmonicity, pulse duration, delayed harmonics, and envelope type. The most important aspect to Patterson's sounds was the use of psychoacoustic cues to indicate urgency (repetition, louder, faster, more abrupt onset for greater urgency), and that the Emergency form was essentially the same as the Caution form, but repeated twice, louder, faster and more intrusive.

The ISO 9703.2 standard published in 1994 [14], described a sound set that was basically a simplified version of Patterson's "general" sounds. It did not make use of any of Patterson's psychoacoustic cues, and the underlying sounds, which were based on simple beeps, lacked an information signal and category specific sounds. It did, however, serve as the governing specification for alert sounds, since it provided a model for audible alerts that could be more easily emitted than Patterson's sounds with the available technology. (This is the technology still employed in the critical-care units at the University of Iowa Hospitals and Clinics.)

Only in 2005, when ISO IEC 60601-1-8 was published for international use [15], were any attempts made to improve on the 9703.2 standard. 60601-1-8 brought back an information signal, provided advice on how to implement psychoacoustic cues, and defined a standard for melodies that matched Patterson's seven categories, plus an additional category for powering a device down. 60601-1-8 standard alerts are based on

melodies consisting of sequences of single notes within a single octave (e.g., c d e (pause) f g). Each individual sound pulse must have a fundamental frequency somewhere between 150 to 1000Hz, and there must be at least four harmonic sounds from 300 to 4000Hz.

The 60601-1-8 standard has created much controversy, and studies of the 60601-1-8 melodic sounds have pointed out their inadequacies [2], [16], [17].

High priority sounds closely resemble the medium priority melodies, with two additional beeps after the base melody, so that each medium priority alarm is readily confused with its high priority version and is easier to learn [18]. The resulting range of nearly 13 different sounds is problematic as people typically can only learn 6-8 sounds of similar form [19].

Studies have shown that the alarms are hard to learn and easily confused with one another. Sanderson, Wee and Lacherez have demonstrated the sounds are difficult to learn, favor listeners with musical training, and do not necessarily differentiate between high and medium priority [3], [1]. Block, a member of the ISO committee that designed the 60601-1-8 standard, has conceded that the alarms are not ideal [20] ("We have suffered long enough with poorly-designed alarm sounds. It is time to correct the mistakes of the past and move forward" [20]).

Edworthy suggests that the idea of melodic alarms ignores practically all the research that has been done in the design of alarm sounds [21]–[23]. They are difficult to learn and retain and so are a disadvantage right from start. In 2013, Edworthy built three sets of eight alarms that were mapped to the categories in the 60601-1-8 standard. One set consisted of tonal alarms that met the standard, the second set consisted of abstract alarms randomly selected from a database, and the third set was designed as indirect metaphors of the functions. When participants were presented with the alarms and asked to indentify them across ten block of eight trials, the results indicated a significant difference in learnability between the alarm sets, with the non-standard alarms sets being

easier to learn that the standard, especially the “metaphor” sounds which the participants learned “significantly better” [24].

Research that aims to identify sounds that are candidates for good audio alerts has mostly been performed with experiments that attempt to identify the "semantic differential" between sounds that are intended to suggest a specific meaning or induce specific emotional effects (e.g., a sense of urgency) [25]. The other most prominent theme in the research corpus has been retention and learnability.

2.2.2.2 Semantic Differentiation.

In *" On the Sensations of Tone as a Physiological Basis for the Theory or Music"* [26], Helmholtz provides an early example of semantically mapping timbre as a sound attribute, concluding that “the quality of the musical portion of a compound tone depends solely on the number and relative strength of its partial simple tones and in no respect on their difference in phase”. In the 1970s, Von Bismarck [27] and Pratt et al. [28] both found that construction of a verbal subjective rating scale for the timbre of sounds is feasible.

The more recent work of several researchers in both healthcare and non-healthcare settings has demonstrated effective models of timbre parameterization for consistently eliciting sense of urgency. In experimental settings that used steady-state, broadband, ambient noise (Haas et al 1995 [29] and Hass and Edworthy 1996 [30]), alerts suggesting increased urgency – typically loud, high-pitched sounds that repeat often in a given time interval- elicit responses with shorter reaction times in subjects [29]–[32].

Burt et al. [32] found physiological responses, including event-related potentials on the electroencephalogram (EEG) to correlate with decreased reaction time in response to high-urgency alerts.

However, perceived urgency is not always the best indicator of the effectiveness of a sound as an audio alert.

As pointed out in [29], [30], [32], [33], there is often a serious mismatch between an alert's *perceived* (psycho- acoustic) urgency, which is a function of its sound parameters, and its *situational* urgency, which is the degree of urgency that the experienced operator (e.g., pilot, nurse) has learned to associate with the warning as a function of the situation itself.

Loud, high-pitched, insistently repeating sounds, while appropriate to *draw* the operator's attention and to be *noticed* over ambient noise, are not necessarily most conducive to efficient operator functioning in *addressing* the indicated problem, especially if the corrective action involves a complex series of steps that require multi-tasking and mental concentration (e.g., arresting ventricular fibrillation or cardiac arrest). Most individuals function best within a narrow range of stress, and anxiety-inducing, insistent repeated alerts then become a background distraction that the operator is required to mentally tune out in order to focus on the problem. The level of distraction can be significant when the operator is trying to process numerous simultaneous inputs. Some degree of operator control over alert parameters after the initial alert has sounded is therefore highly desirable.

As pointed out by Kerr et al [4], such tones have the undesired side-effect of inducing anxiety in a patient who is rarely in a position to do anything about the problem. As this thesis discusses later, modern technology is readily capable of addressing this issue.

2.2.2.3 Recognition and Retention

Recognition and retention of audible alarms is important. Milligan et al. found that nursing staff could only recognize 83% of their alarms (compared to 43% of medical staff, who are in the unit for significantly less time per day than nurses), and that the three most critical alarms were recognized poorly [19]. Williams and Beatty, using 21 non-clinician subjects to identify IEC 60601-1-8 alarms melodic alarms as defined by Block et al. [17], found that identification accuracy was only between 10% and 61% and the

participants often confused alarms . Sanderson & Wee had 22 critical care nurses learn the IEC 60601-1-8 alarms over two training sessions more than a week apart and found an average recall rate of 66%, with only one nurse having 100% accuracy. They concluded that “IEC 60601-1-8 melodic alarms should be redesigned before they are adopted for clinical practice”[2].

2.3 Limitations of Past Work

The premise of this thesis is that:

1. Many of the efforts made toward devising "standard" alert sounds, while possibly relevant pre-2000 because of limitations in the ability of devices to store and reproduce complex sounds, may be much less relevant in today's era of affordable, high-powered, wireless-network-capable devices. Several previous research efforts, including many models for audio alert design, have been encumbered by implicit assumptions about hardware capabilities and technological stasis that are not valid anymore. Consider, for example, that hand-held devices like the iPhone and Android family have more computing power and memory than the world's most powerful supercomputers of the mid 1980s, and that they offer ubiquitous connectivity with other devices.
2. Standardization is not necessarily compatible with rapidity of learning and accuracy of retention, especially across cultures and between individuals. For example, while people with musical ability are able to learn and differentiate between 60601-1-8-standard alerts more readily [3], one might expect that with particular individuals, specific samples of genres of music might be more memorable than others, and that non-musically-trained individuals might choose non-musical cues instead. This point is discussed in the Results chapter when we discuss the diverse population strategies employed by the users that we tested.
3. As long as one achieves quick learning and high retention, it is not even certain that the search for the perfect, "standard" set of sounds is meaningful, any more than the

search for a user-interface approach for software or hardware design that never needs customization or adaptation. Many implicit assumptions in research of universal cultural norms may be questionable. For example, there is evidence of cultural drift with respect to sense of urgency response to pulse based tones [31]: sequences or combinations of tones that a previous generation may have found alarming/discordant may not arouse the same emotional responses today.

4. The trend in technology has been to progressively reduce the user's cognitive burden so as to minimize the possibility of human error, and personalization is one way to achieve this goal. A widely used textbook on software-user-interface design by Steve Krug is appropriately titled, "Do not Make me Think" [34]. Examples of technology that operates in this fashion are cited below.

    a. When numerous signals are generated near-simultaneously, there is a significant cognitive burden involved in mentally translating the corresponding alert sounds into their meaning, and then performing an action based on that meaning. In fields such as Formula One car racing, where alerts to drivers have long been personalized, a driver must track a very large number of simultaneous inputs: the driver's crew helps to minimize the driver's cognitive burden by monitoring the status of the car remotely and providing direct verbal input to the driver.

    b. In signal transmission, the historical dependence on Morse code, which can work in the most primitive and noisy conditions, but which requires significant operator skill and training, was largely eliminated as electronic transmission became a more reliable and error-recovery protocols became more robust.

    c. High-end automobiles use speech synthesis to directly prompt the driver to perform an action, as opposed to merely emitting beeps or alarms that the driver must interpret. GPS Navigation aids employ speech synthesis to provide directions.

In summary, alarm design should take into account the opportunities for personalization that technology provides. The role of personalized alerts is now

envisaged in a futuristic scenario that is, however, well within the scope of today's technology.

1. When a healthcare provider starts their shift at the ICU and logs in, a central server looks up the provider's personalized alert sounds and transfers them to the bedside devices that are connected to the patients that the provider is responsible for during the shift. The peripheral device also possesses the ability to record sounds much like today's smartphones.
2. The provider uses a mobile earpiece, to which the alert sounds are transmitted wirelessly from the bedside device. The personalized sounds, which are meaningful only to a particular provider, are therefore heard only by that provider rather than by the patient who is in no position to do anything about them. The use of earpieces will address a major concern about audible alarms today: being inundated with their own alarm sounds as well as everyone else's makes it very difficult for patients in ICUs to get any rest.
3. The bedside device may multi-cast, so that if more than one provider is responsible for that patient, each provider will hear different sounds. With proximity sensing, high-priority sounds may be sent to the physically most proximate provider even if that provider has not previously been designated to manage that particular patient.)
4. In emergency situations, some patient-management chores that do not reqire nursing expertise– e.g., rotating the patient to prevent pressure ulcers, or to assist respiratory drainage – may make sense to delegate to volunteers or relatives. Unskilled volunteers cannot be expected to recognize the full range of alerting sounds. Instead, the range of alerts may be customized to allow them to recognize a narrower range – a set of sounds that they can act upon with respect to performing specific actions on the patient, and a "catch-all" sound that simply indicates that they should seek the help of a nurse.

5. The alert sounds need not necessarily be beeps or musical tones. A short spoken phrase stating the problem (or the corrective action to be performed) might in fact be preferred, since it allows bypass of the step of mental translation of the sound into the action.

Further, unlike speech synthesis or pre-recorded messages that are recorded in a particular language, personalized messages can be recorded in the user's own language, and even the user's own voice. This makes them applicable to situations where clinical monitoring equipment designed in one country is deployed elsewhere (e.g., In the emergency situation, in a third-world country during international aid efforts for a natural disaster), customized alerts could be phrases in a non-English language that are recorded on location.

CHAPTER 3 METHODOLOGY

This chapter describes the tools and methodology used to create our study, and the results and observations that the study provided. First we'll discuss the design criteria for a personalized alerts study platform in section 3.1. Section 3.2 will introduce the custom application that was built to meet that criteria; PASA. Section 3.3 will describe the study that was designed and executed with the application. The study results will be presented in section 3.4, and in section 3.5 we'll summarize and discuss those results.

3.1 Application design criteria

With the exception of some studies that focused solely on measuring semantic differentiation, all of the previous experiments studying audible alerts utilized an integration of software and hardware that sounded alerts and recorded information of interest about a listener's response to those alerts- usually recognition, retention, and response time, or some combination thereof. Typically the platform developed for the study was constrained to the specific parameters of the individual study and therefore was a "one off" and not intended to facilitate future research.

It quickly became apparent that studying these responses with respect to personalized alerts required a platform that provided capabilities not previously considered or necessary. For instance, it was necessary to create a simple means for capturing the sounds that a user might want to use for their personal alerts, one that could readily accommodate the variety of sources that a study participant might draw upon for sounds such as CDs, ringtones, voice, video, etc.

Once the sound was captured, the ability to edit clips consistently in an expeditious manner was also necessary in order to best facilitate subject participation in the creation of their personalized alerts. The sound had to be visible in a waveform editor so that cropping, normalization, and Hanning windowing [35] functions could be easily applied so as to homogenize gain levels and eliminate the annoying clicks and pops that

usually result from cutting pieces of audio out of a longer selection. Several DAW (Digital Audio workstation) applications were considered[1], but they were deemed too cumbersome to allow timely collaboration between the researcher and study participant.

Additionally, once an alert was created, there had to be a means to accommodate the participant's sense of repetition, i.e., what type of looping and interburst attributes would make the sound more effectively convey the priority level that they were mapping it to.

Because all of the alert attributes had to be persisted across test sessions, a data store that would respond to requests quickly enough to allow for individual sounds to be loaded into playback buffers in real time would have to be designed. Each sound needed to be indexed and recorded, along with volume information, looping parameters, and information the application needed to afford playback, such as sound type and clip length.

All of the above requirements needed to be met not only for personalized sounds, but for any type of sound that a study might be designed to contrast against, e.g., sounds meeting different standards, abstract sounds, earcons, or simple voice commands.

Since personalized alerts had not been studied before, it was also determined that building an interface that would allow study parameterization to be very flexible was also needed. All previous research on retention and recognition had been done with predetermined sounds, and few of the research articles provided additional insight by running successive sessions with changed control parameters. We wanted our tool to function as a protocol designer that would serve as a hub not only for current research, but for future research on personalized information in the context of health care.

---

[1] Logic Pro, Ableton Live, Audacity, Protools, Garageband, Soundtrack Pro.

The interface that a study participant uses to make selections during the execution of a study is also important. The interface must be simple to understand, easy to use, and offer a means to present any desired feedback to the study participant during the course of a test session. These attributes provide interface credibility [36], and in doing so, minimize bias that the interface might affect upon study results.

The level of dynamic malleability that the touchscreen based interface provides makes it an ideal platform upon which direct manipulation interfaces can be developed. It is a device that has become highly ubiquitous and is easy to learn. It can easily host interface elements, such as buttons for instance- that dynamically change as required by parameters of the application, e.g., button location, number of buttons, and images within the buttons (icons). Touchscreen buttons, often referred to as "soft buttons", offer performance that is similar, and in some cases, even superior to hard buttons [37].

3.2 PASA (Personalized Alerts Study Application)

Once the design criteria was established, an appropriate combination of hardware and software was selected that could be used as a development platform for what would eventually become PASA, short for Personalized Alerts Study Application. PASA can be run on either a Windows PC or a Macintosh. The touchscreen selected was Apple's iPad. PASA communicates with an iPad through a wireless connection and PASA controls the interface that the study participant interacts with on the iPad. PASA also has the ability to synchronously control multiple iPads. This feature allows a study protocol to be designed that could host several participants reacting to personalized alerts simultaneously. See figure 1 for a screen shot of PASA and figure 2 for an example of an iPad screen rendered and controlled wirelessly by PASA.

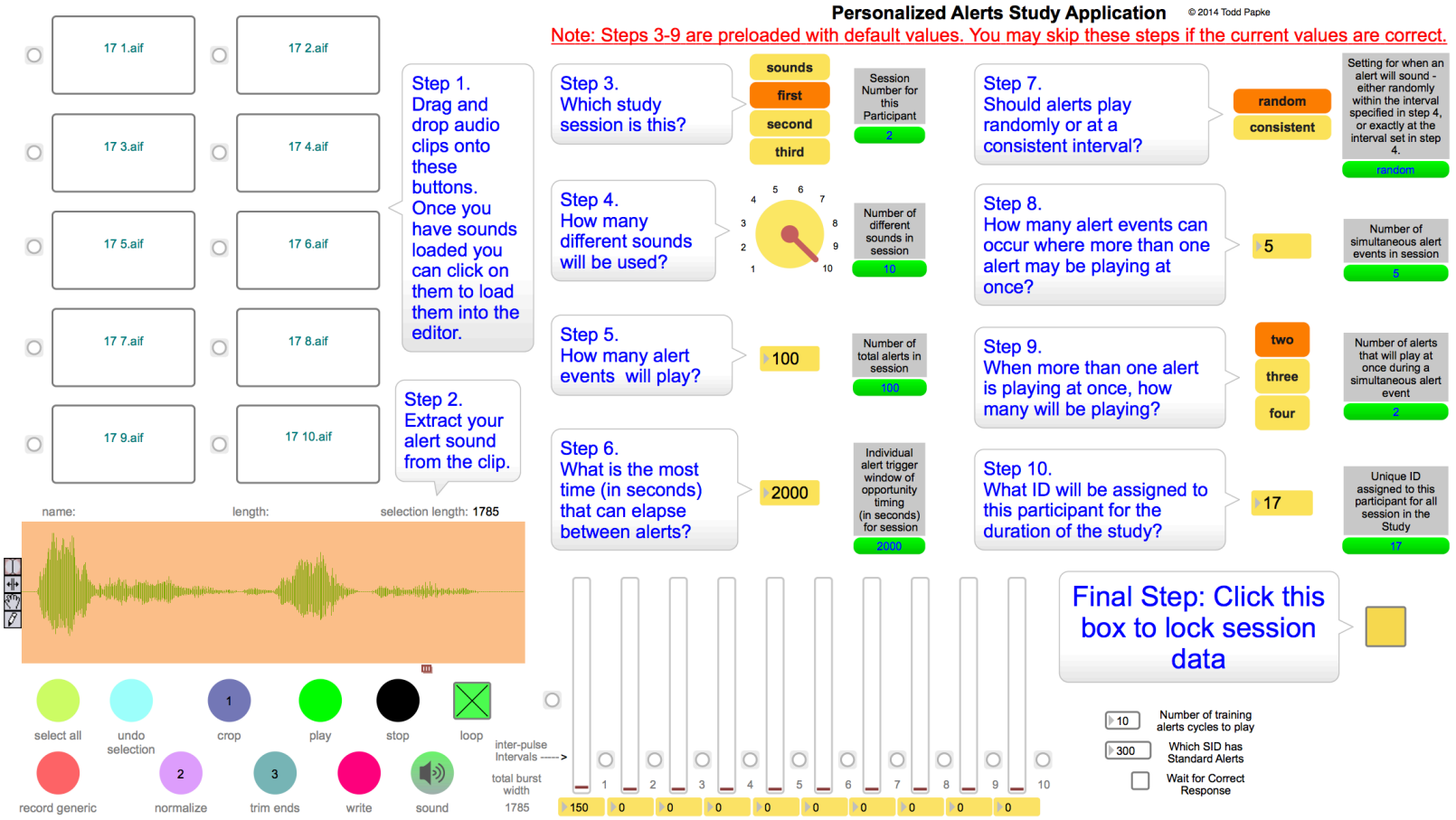

Figure 1. PASA interface

PASA was developed with Max/MSP. The iPad interface is rendered by MIRA. Max/MSP and MIRA are both development tools available form Cycling ’74 Inc. [38]. Applications developed with these tools can run unmodified on both Macintosh and Windows computers. Max offers similar capabilities as Csound, the public domain application used by Wee and Sanderson to develop a study on melodic alarms in 2008 [2], but adds a visual programming interface that combines elements of both traditional

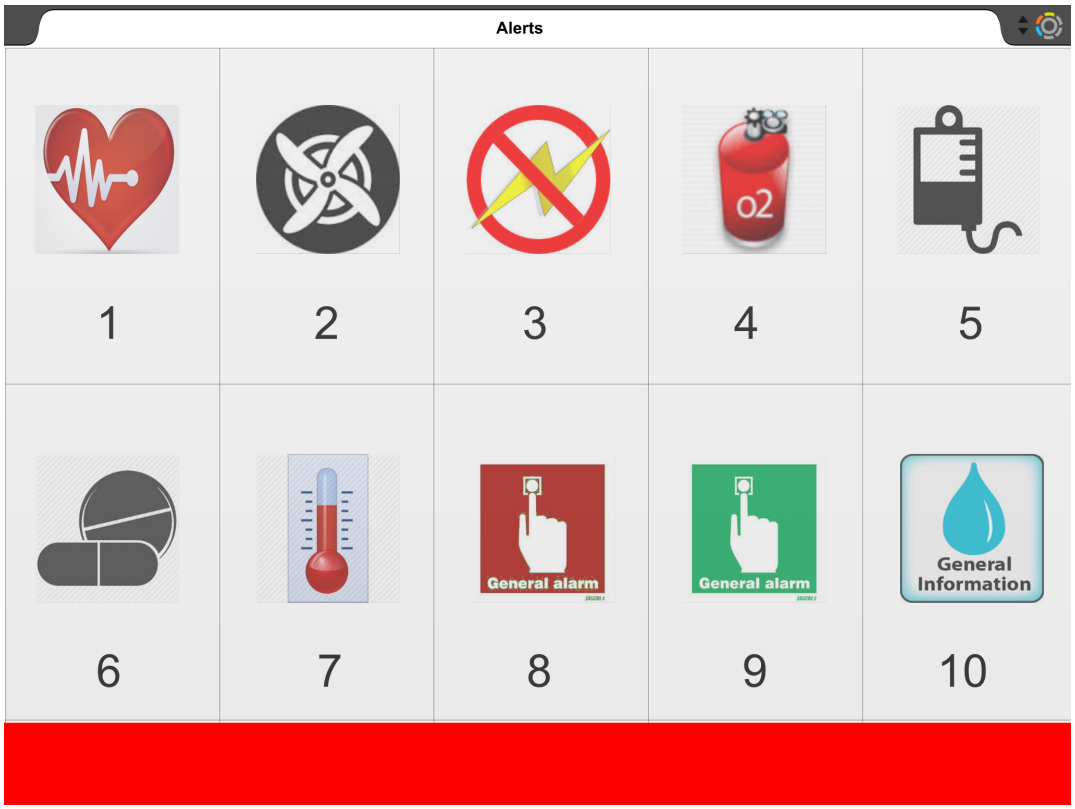

Figure 2. iPad interface for 10 alerts

programming languages and circuit design. It is ideal for rapid prototyping, and its native programming environment can be augmented with libraries written in both Javascript and JAVA.

MIRA is an application that runs natively on the iPad. It allows a Max application to render and draw an interface, and control that interface in real time through a wireless connection. This offers several advantages over the traditional iOS development environment. Applications can be developed that are not constrained to OS X, the development process does not require time consuming compile and load cycles, and the Max application can serve as the central hub for all interface and logic processing, which eliminates the need to develop separate applications for both computer (server) and touchpad (client). For a screen shot of the iPad interface that was created with PASA and rendered with MIRA for the 10 alert study, see figure 2.

PASA allows sound to be extracted from an AIFF audio file and trimmed to be effectively used as an audio alert. PASA has a waveform display panel that allows the sound clip to be edited visually (i.e. cropped, normalized, trimmed), and can store up to ten sounds for each study participant. Each sound, corresponding to an alert, has a "slot" in the study design interface (fig 1., top left zone) for selection and manipulation. PASA displays clip length, so a sound clip can also be created that meets any sound pulse length desired. Additionally, PASA has ten inter-pulse delay controls, plus an inter-burst delay control. This allows sounds to be created that comply with the inter-pulse and inter-burst requirements of both the ISO 9703.2 and ISO 60601-1-8 standards. Sound files can be dragged and dropped into the application, recorded into sound bank slots directly, or recorded as simple named files that can be edited and assigned to banks at a later time.

As a tool that facilitates designing study protocols, PASA can be used to create banks of personalized sounds and run a study that tracks response time, retention ability and recognition between any two banks of sounds. In our case, the personalized sounds were contrasted with sounds that meet the most current ISO standard, ISO 60601-1-8. Study participants are keyed with a numerical ID, which the study designer can specify. In design mode, PASA will alert the study designer if they are attempting to modify a subject ID that has already been assigned, and suggests a study participant ID that has not

been used (see figure 11). In session mode, PASA will automatically begin and end a study, prompting the study participant when new sections of the study are to begin. For instance, in our first study session, there were four sections, (1) sound bank 1 training section, (2) sound bank 1 testing section, (3) sound bank 2 training section, and (4) sound bank 2 testing section. In the follow-up session there were only two: bank 1 testing section and bank 2 testing section. (see section 3.3).

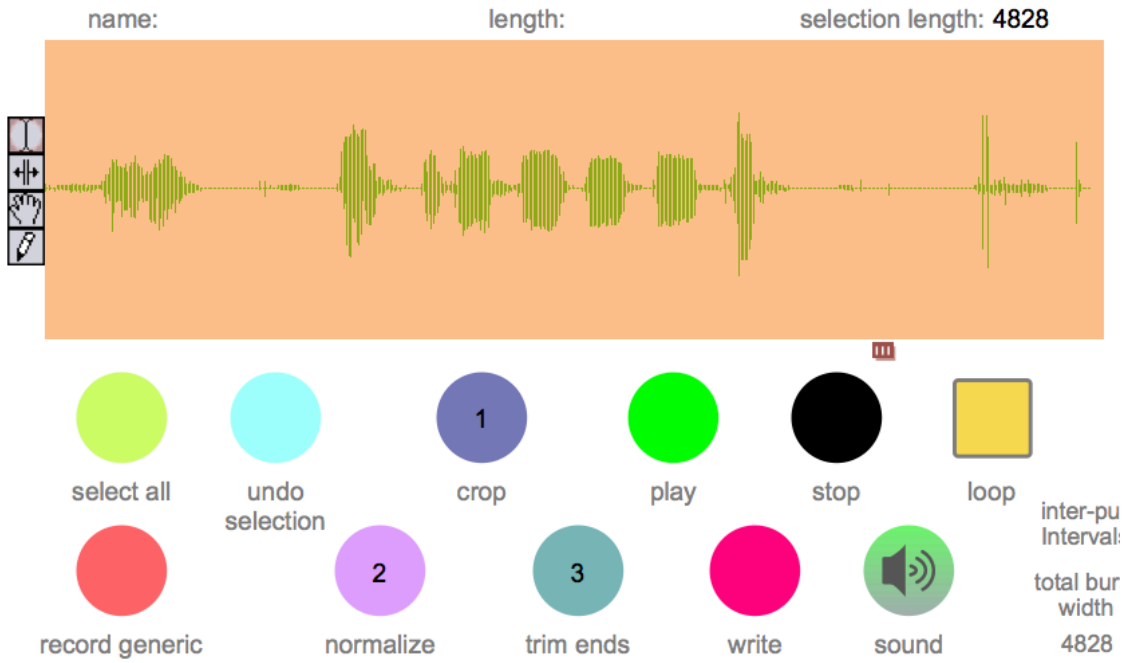

Figure 3. PASA waveform editor

3.2.1 PASA audio capture and editing capabilities.

The audio editing section of PASA is designed to allow comfortable interaction between the study participant and the researcher by providing a means to easily record a sound and quickly edit that sound to make an acceptable alert. Figure 3 shows the visual waveform editor that allows the researcher to quickly zero in on the specific section of an audio clip and respond to the study participant's change requests interactively. This

visualization of the audio clip provides a simple, self-explanatory means to convey what each edit is doing to the sound, thus facilitating deeper collaboration between the researcher and study participant. In our study, all study participants, after seeing the results of an edit sub-step and listening to the result, immediately understood what was being done to their initial "raw" sounds and enthusiastically contributed to the editing process. This gave them confidence that they had indeed "personalized" their alerts.

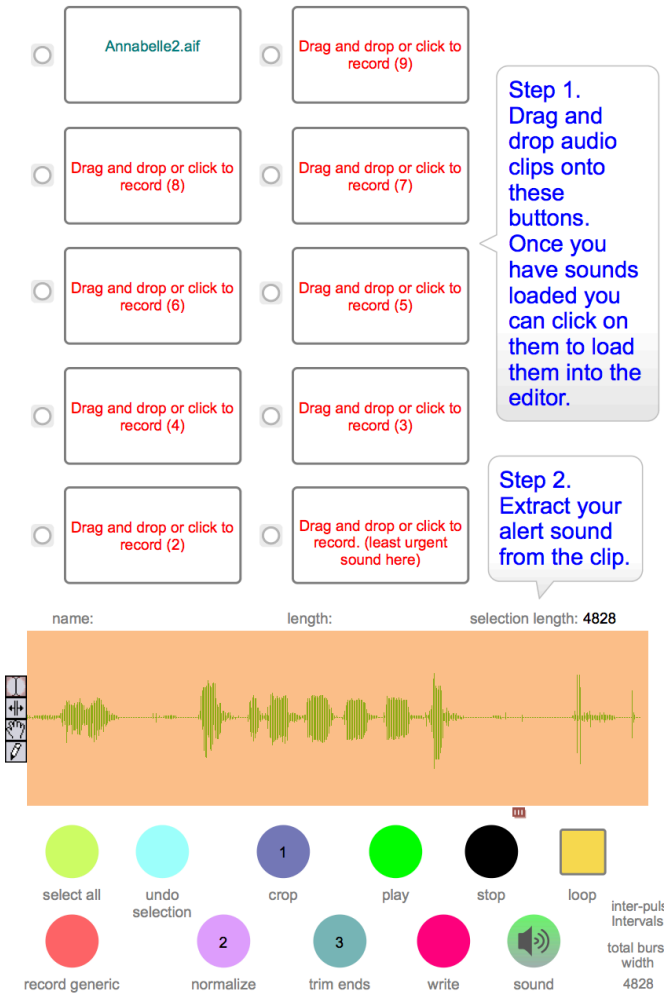

Figure 4. PASA alert sound slots

3.2.1.1 Adding Sounds to PASA

Figure 4 shows the sound slots in PASA. To begin the process of creating an audio alerts PASA provides a number of starting points. (1) A Sound can be added with 'Drag and Drop'- by dragging the file from a directory window into the alert slot desired, (2) An alert sound can be recorded into a slot by clicking on an empty slot. The slot will go into record mode and begin recording. A second click on the slot will stop the recording and the recording will be immediately loaded into the waveform editor. (3) The 'record generic' can be used to navigate to an area in the file directory and give a recording a name. This can be useful if several recordings need to be made, and the

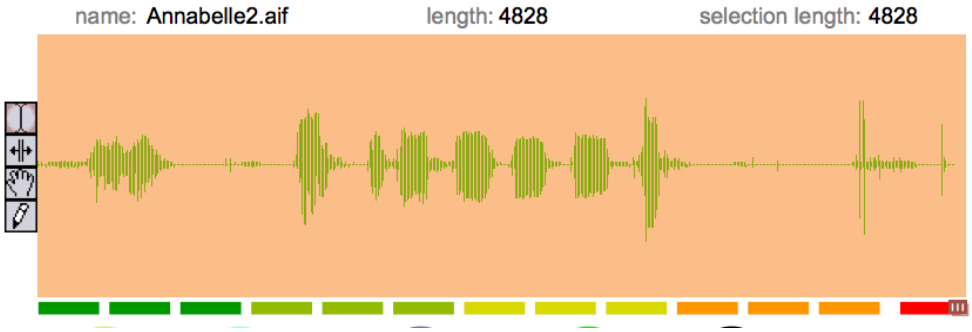

Figure 5. PASA waveform editor

sounds are to be dragged into slots at a later time. The 'stop' button will end a generic recording. Gain controls in slot recording and generic recording work the same, and the level is shown below the waveform editor (see figure 5).

3.2.1.2 Editing Sound Clips in PASA

Once a sound is recorded or loaded into a slot, clicking on the button will load the waveform editor with the sound. At this point the sound can be played once by clicking on 'play'. Turning on looping will allow the sound to loop, and the looping characteristics can be set-up with the loop controls. While looping, the sound can be

edited with the editing controls, and, as the individual edit function is applied, the waveform window will show the edit effect on the sound, and the sound will audibly change to reflect the effect. This allows for 'real time' editing; the result of an edit function is immediately reflected visually and audibly. The suggested steps for editing

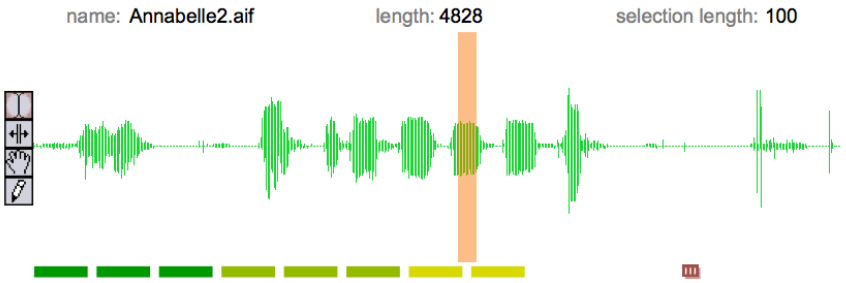

Figure 6. Audio clip selection

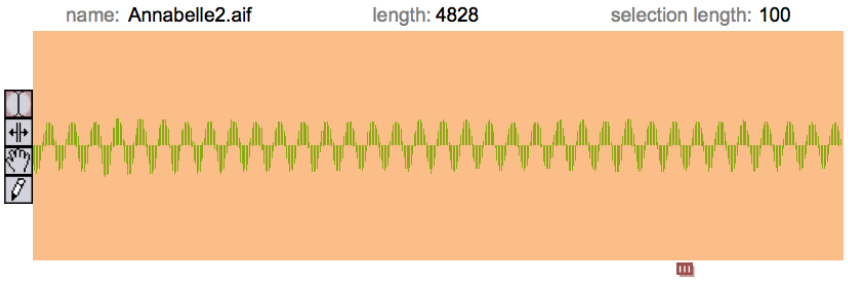

Figure 7. Selection cropped

are:

1. Select the section of the audio clip that is desired for the alert sound and use the cropping tool. The selection function automatically floats the ends of the selection to zero crossing points in the audio clip. This is done to eliminate clicks that would otherwise occur. If looping is turned on, the selection will play in a continuous loop, which gives constant feedback of the selected audio, allowing the precise section of a clip

to be identified audibly when selecting. The selection palette on the left side of the waveform editing window provides various types of selection (see figure 6). From top to bottom:

a. I-beam icon: Click at a location and drag.

b. Double beam icon with arrows pointing both ways: Click on a location and move up or down to expand selection equally to the left and right of the selection point.

c. Hand icon: Zoom in on waveform.

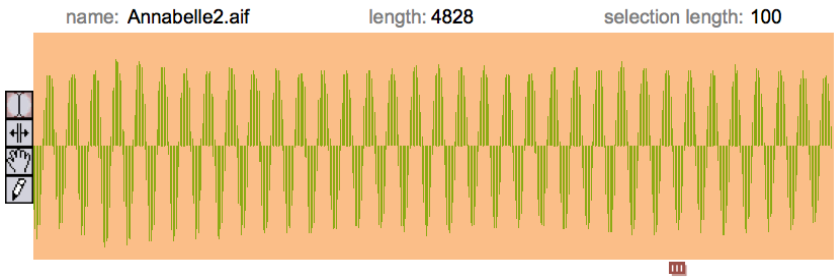

Figure 8. Selection normalized

d. Pencil icon: Draw on waveform (currently disabled).

2. Crop the selection. When section of the clip that will become the alert sound is selected, the ‘crop’ button will cut the selected section and fill the entire waveform editing window with it. This is the first step in editing and the crop button is labeled with the number ‘1’. See figure 7.

3. Normalize the selection. This is a standard audio normalization function. A constant amount of gain is applied to the clip to bring the peak amplitude to a target level. This function works very well when the recording gain is too low for the gain control to increase amplitude to an acceptable level. The normalization button is labeled with the number ‘2’. See figure 8.

4. Trim the selection. This function uses a Hanning [35] window function to apply a waveform smoothing effect to the beginning and end of the selection. Smoothing parameters have been selected that result in a tone that slowly increases in amplitude as the tone starts, and gently decreases in amplitude as the tone ends. Often the normalized sound is appropriate for an alert, but if there is a desire to give the alert a ramp up and ramp down to increase effectiveness, the trim function can be very effective. It also works well when the clip has an aggressive amplitude ramp, as this often can cause a pop. Trimming is designated by the 'trim ends' button which is label with number '3'. See figure 9.

While the editing steps outlined above were designed to produce optimal results, the editing functions can be done in different orders than those outlined above. Practicing editing functions on audio clips will provide some insight into which edit functions work best with which type of audio clips. There is no undo, however, clicking on the sound bank button will reload the original clip for editing, allowing editing to be started over. When the study participant is pleased with the alert, the 'write' button will commit the edited alert to the sound bank.

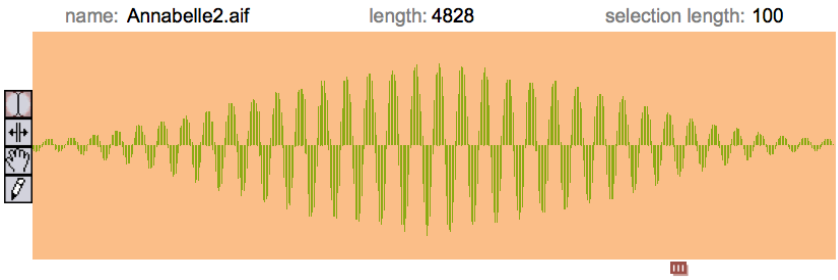

Figure 9. Selection trimmed

Table 1. ISO 60601-1-8:2012

Characteristic	HIGH PRIORITY ALARM SIGNAL	MEDIUM PRIORITY ALARM SIGNAL	LOW PRIORITY ALARM SIGNAL[d]
Number of PULSES in BURST [a, e]	10	3	1 or 2
PULSE spacing (t_s) (see Figure 1)			
between 1st and 2nd PULSE	x	y	y
between 2nd and 3rd PULSE	x	y	Not applicable
between 3rd and 4th PULSE	$2x + t_d$	Not applicable	Not applicable
between 4th and 5th PULSE	x	Not applicable	Not applicable
between 5th and 6th PULSE	0,35 s to 1,30 s	Not applicable	Not applicable
between 6th and 7th PULSE	x	Not applicable	Not applicable
between 7th and 8th PULSE	x	Not applicable	Not applicable
between 8th and 9th PULSE	$2x + t_d$	Not applicable	Not applicable
between 9th and 10th PULSE	x	Not applicable	Not applicable
INTERBURST INTERVAL [b, c] (t_b)	2,5 s to 15,0 s	2,5 s to 30,0 s	>15 s or no repeat
Difference in amplitude between any two PULSES	Maximum 10 dB	Maximum 10 dB	Maximum 10 dB

Where: x shall be a value between 50 ms and 125 ms,

y shall be a value between 125 ms and 250 ms,

the variation of x and y within a BURST shall ~~be~~ not exceed ± 5 %, and

MEDIUM PRIORITY $t_d + y$ shall be greater than or equal to HIGH PRIORITY $t_d + x$.

The INTERBURST INTERVAL (t_b) for HIGH PRIORITY auditory ALARM SIGNALS shall not be greater than the INTERBURST INTERVAL for MEDIUM PRIORITY auditory ALARM SIGNALS which shall not be greater than the INTERBURST INTERVAL for LOW PRIORITY auditory ALARM SIGNALS.

[a] See also Table 4 for characteristics of the PULSE.

[b] Unless otherwise specified in a particular standard for a particular ME EQUIPMENT.

[c] MANUFACTURERS are encouraged to use the longest INTERBURST INTERVAL consistent with the RISK ANALYSIS. Writers of particular standards are encouraged to consider the longest appropriate INTERBURST INTERVAL of the auditory ALARM SIGNAL for the particular ALARM SYSTEM application. Long INTERBURST INTERVALS can under certain conditions negatively affect the ability to correctly discern, in a timely manner, the source of the ALARM CONDITION.

[d] The generation of the auditory component of a LOW PRIORITY ALARM CONDITION is optional.

[e] Unless inactivated by the OPERATOR, MEDIUM PRIORITY and LOW PRIORITY auditory ALARM SIGNALS shall complete at least one BURST, and HIGH PRIORITY auditory ALARM SIGNALS shall complete at least half of one BURST.

Source: International Standards Organization, Geneva, Switzerland, IEC International Standard IEC 60601 (2012): Medical Electrical Equipment – Part 1–8: General Requirements for Safety. International Standards Organization, Geneva, Switzerland.

3.2.1.3 Looping Alerts in PASA

PASA contains a loop controller that allows inter-burst timing. This is the first slider in Figure 10 below. Used alone, this slider will allow the delay between an alert to be specified. Additionally, PASA has ten inter-pulse interval sliders. These can be used to build a series of delays between alert sounds. With these controls, an alarm can be created that meets the tone burst requirements of the 60601-1-8:2012 standard (See Table

1). Figure 11 shows an alert with pulse and burst parameters set to meet the requirements for a high-priority alert as specified in table 1, which is taken from the ISO 60601-1-8 specification.

To empty a slot, or set the looping controls to zero, there are radio buttons by each

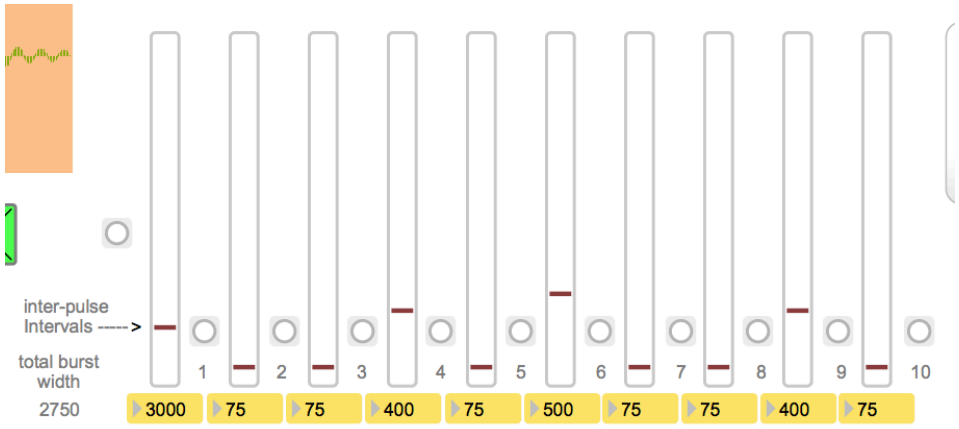

Figure 10. Alert pulse and burst parameters that meet 60601-1-8:2012 requirements

control. In the case of sound slots, clicking on the radio button will empty the sound slot so that it can be re-recorded, and clicking on the radio buttons by each loop control will zero the control. Also the radio buttons by each loop control will illuminate while the sound is playing to indicate where in the inter-burst sequence the sound is playing. This gives a visual indicator that corresponds to the sound loop characteristics. Much like the waveform editor visual feedback, this gives visual feedback in an effort to understand what the loop controls are doing with respect to the alert sound characteristics.

The gain (volume) and looping parameters are stored on a sound-by-sound basis. When a sound bank is loaded, each sound has a corresponding parameters record that is loaded with each sound into the playback engine during testing, or recall. When all of the alert sounds have been recorded and edited, clicking on the toggle box will lock the

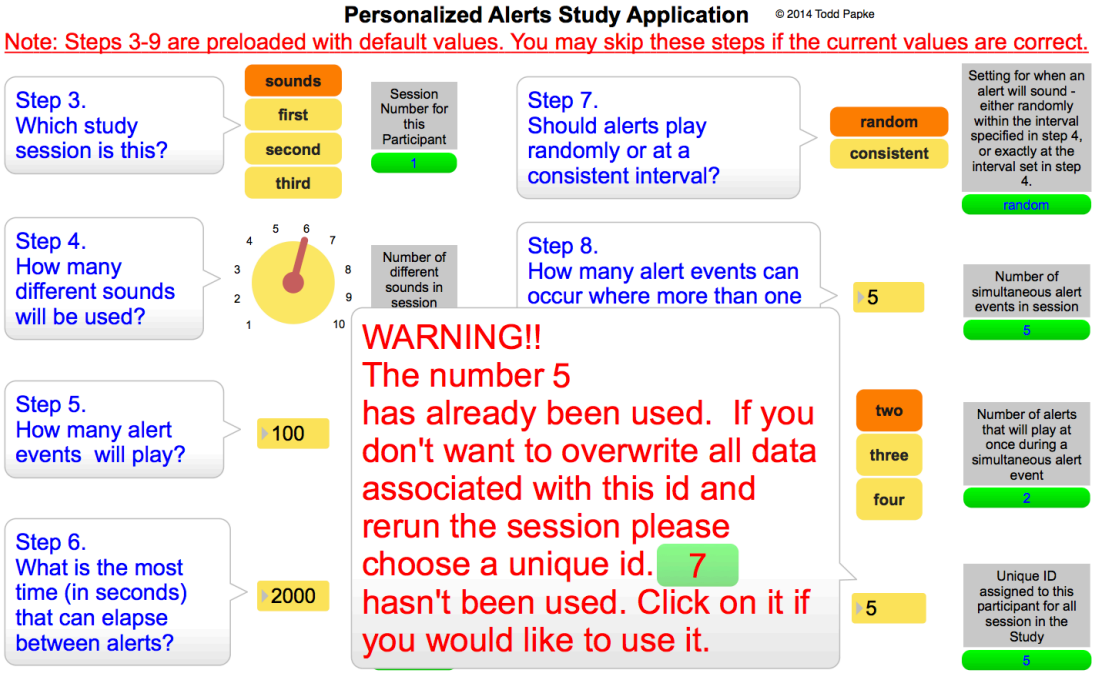

Figure 11. PASA study design control panel

parameters, and, depending on the session selected either write out all information or begin a study (see figure 12 and figure 13).

3.2.2 Designing a Study with PASA

PASA's study design control panel allows a researcher to customize their study protocol by providing a means to manipulate several study parameters (see figure 11):

1. Total number of alerts to be played to the study participant for each study section. The control in figure 11 is set at 100 alerts, which means that the subject will respond to 100 alert events for their personalized sounds and 100 alert events for the sounds loaded into another sound bank. In our study, the second bank contains sounds that meet ISO 60601-1-8 requirements.
2. Total number of different alert sounds to be used in the study (1-10). The control in figure 11 is set to use 6 different sounds.

3. Interval of time in which an alert can play specified in milliseconds. For instance, 2000 milliseconds means that one alert will always play within two seconds after a previous alert has been responded to, unless a test session is just beginning and the alert being played is the first alert. The parameter in figure 11 is set for 2000 milliseconds.
4. Random alert setting. This indicates whether alerts should play after the timing interval has passed or randomly within the interval specified. The option selected in figure 11 is for random alerts.

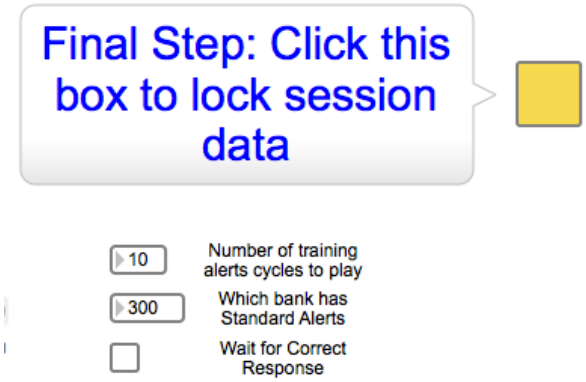

Figure 12. Session lock, training cycles and comparison bank

5. If a study designer wants study more realistic alert scenarios, the study can also include events where more than one alert plays within the specified timing interval. In figure 11 the parameter is set to play simultaneous alerts 5 times. These events will be evenly distributed within the 100 alerts, so a simultaneous alert event will occur every 20 alerts (100/5 = 20). Also the number of simultaneous alerts can be specified – in figure 11 the number of simultaneous alerts played will be 2. Note: simultaneous alerts will occur randomly within the

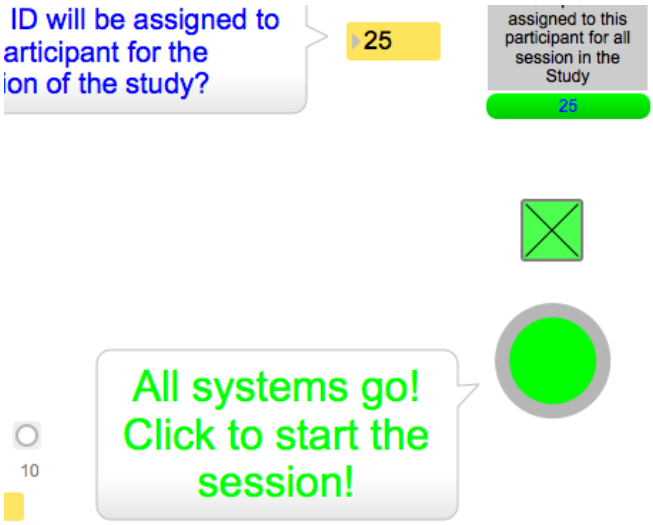

Figure 13. Session write and study execute

alert time interval (3 above). This allows for a more realistic experience; alerts could occur at the same time, or stagger randomly.

6. Different aspects of a study are executed with the session selector. In figure 11 the selection is “Sounds”, which indicates that PASA is in sound capture and design mode for the particular bank; “first” indicates that you want to do a study that begins with sound training, “second” and “third” are sessions in which a test can be conducted without training.
7. Alert training cycles: number of times the application cycles through the alerts during the training phase of a test sequence. In figure 12, the study is set to cycle the alerts 10 times.
8. Alert Bank for comparison: which bank contains the alerts that are to be included in the test with the personalized alerts. In figure 12, the study is set to compare the personalized alerts with the alerts in sound bank 300.

9. Wait for Correct response: When this is checked the alert will continue to play until the study participant touches the correct response. In figure 12, the study is set to have a new alert play whenever a panel is touched.
10. ID overwrite warning: when the study session is set to ‘sounds’, PASA is in alert design mode. When a study participant ID is entered into the ID box, PASA will warn that the ID may be overwritten if there is already alert data present for that participant ID. In figure 11, PASA is warning that participant ID 5 is already in use, and that participant ID 7 is available. Clicking on the green button within the dialog will automatically assign 7 as the participant ID.
11. When the session information is correct, the check box in figure 12 will lock the session data. At this point the check box will turn green and the commit button, ‘All systems go!’ will be displayed. Clicking on this button will write out all study session data if PASA is in design mode (session ‘sounds’). For any other mode, clicking the green button will begin the study. Session 1 will run the study with training sections, session 2 and 3 will run a study with just the testing cycles.

3.2.3 PASA data collection and error recovery

PASA automatically builds a persistent data table with records that include the following information. The index key into the table is the study participant’s ID. See Table 2.

During the execution of a study, the following attributes are recorded:

1. Which Section of the study does the record refer to, i.e., training for the personalized sounds, testing for the personal sounds, training for the contrasted sounds, or testing for the contrasted sounds. In Table 2, the section shown is ‘PAT’ or personalized alert training.
2. Target Alert: which alert was actually played
3. Correct Response (1 = yes, 0 = no): whether the study participant selected the correct alert button.

4. Actual Alert Hit: which alert was selected.
5. Total Time (seconds): this is calculated as a function of Alert Start Time – Alert Stop Time; the result is recorded as seconds, with a 1/1000 second accuracy.
6. Alert Start Time (as CPU clock timestamp): this is the time that the alert was played.
7. Alert Stop Time (as CPU clock timestamp): this is the time that the alert was stopped by the study participant making a selection.
8. Participant ID: which participant does the record belong to.
9. Simultaneous Alert (1 = yes, 0 = no): whether the alert was part of a simultaneous alert event.
10. Which Study Session: 1, 2 or 3.

Table 2. Data format key

1	2	3	4	5	6	7	8	9	10
PAT	1	1	1	1.746	949087.125	950833.3125	1	0	1
PAT	2	1	2	2.556	952805.6875	955361.125	1	0	1
PAT	3	1	3	2.183	955708.875	957891.75	1	0	1
PAT	4	1	4	3.466	958655.125	962121	1	0	1
PAT	5	1	5	2.473	963925.125	966398.4375	1	0	1
PAT	6	1	6	2.407	966760.5	969167.375	1	0	1

Study start and stop times are recorded in a separate table (see table 3):

1. Section.
2. Session Start Time (as CPU timestamp).
3. Participant ID.
4. Section.
5. Session Stop Time (as CPU timestamp).
6. Section.

7. Total Session Time: hours.
8. Total Session Time: minutes.
9. Total Session Time: seconds.
10. Which Session.

Table 3. Session timing table.

1	2	3	4	5	6	7	8	9	10
PAT	949040.375	1	PAT	1024162.625	PAT	0	1	15.122	1
PA	1034160.625	1	PA	1393616.625	PA	0	5	59.456001	1
SAT	1403612.75	1	SAT	1497357.625	SAT	0	1	33.744999	1
SA	1507354.625	1	SA	1930177.75	SA	0	7	2.823	1

The set of personalized alerts are stored within the server's host OS file directory system for reuse in future study sessions. Once the alert has been created and edited, a new sound file is written which is renamed [<StudyParicipantID>space<SoundNumber>.aiff]. So, for instance, if the Study Participant ID was 7 and the alert being created was alert 1 (the highest priority alert), the resulting filename would be "7 1.aiff". This is done to allow for alert banks to be automatically reloaded under program control for study protocols in which follow-up study sessions have been designed, as the application is capable of running study sessions automatically.

As a fail-safe, the program is designed to automatically resume the study after a period of 20 seconds if, for some reason, the system becomes temporarily unresponsive to the participant. In such a case, the data point affected by the delay is discarded.

All data is tabulated in a table form that easily imports into tools that might be used for data analysis (e.g, Excel, Access, Minitab, R, SPSS, etc.).

3.3 Experimental Protocol

3.3.1 Hardware configuration

The study was conducted in a 12’ x 12’ room with no special treatment which would alter the acoustic signature. A table with two chairs was placed in the center of the room, upon which all of the equipment used was permanently placed throughout the duration of the study.

The computer used as a study server was a quad-core 15’ Macbook Pro laptop computer. The audio interface used was a Focusrite Saffire 6 AD/DA convertor, which connects to a computer via a USB interface. The Saffire 6 provides two input channels and 4 output channels.

Since the Sapphire 6 is class compliant it does not require special operating system drivers, and allows audio routing to be controlled with software. PASA takes advantage of this, and automatically switches channel routing when the transition is made between recording and editing. This removed any conditions that might create a feedback loop in the audio chain, and allowed the Saffire input and output gain settings to be permanently set at one level for the entire study. All gain settings for recording and playback were done through PASA’s audio editor. This allowed level controls to be easily manipulated and, most importantly, stored and linked to the individual sound files. Each sound, when played, could have its own volume level- which made it very easy to work with audio sources of varying amplitudes, as it normalized the output volumes of each personalized sound for each study participant and diminished the effect that amplitude might have in recognition and retention across the personalized alerts sound collection as a whole.

The touchscreen device was a 3rd generation iPad, which communicated with the laptop through a wireless 802.11n network via a DHCP IP connection. The iPad was not tethered to a power source. For our purposes, the wireless network was configured to be closed, with only the Macbook Pro (server) and the iPad3 (client) connected to it.

Both the researcher and the study participant used headphones. The researcher used Sennheiser PX-100 headphones, and the study participant used AKG K501 studio monitor class headphones. These headphones were selected because they are open-design. This allowed conversation to take place as necessary during audio editing and testing.

A Shure SM57 microphone was used for all recordings that could not be accommodated with a direct line connection, (e.g., voice, acoustic instruments, etc.). It was permanently plugged into line input 1 of the audio interface and situated in a stand that allowed easy manipulation of location. A standard ¼" audio cable was plugged into input 2 of the audio interface. Audio connection adapters were utilized as needed to allow direct connections when necessary with electronic devices (e.g., smartphones, MP3 players, CD players, etc.). All alerts were mixed mono – the same sound for both ears.

3.3.2 Study Protocol

Once the study received IRB approval, a recruiting effort was undertaken in The College of Nursing at The University of Iowa, where members of the faculty were kind enough to allow recruitment presentations to take place in their classes.

Students signed up at the presentations, and were also provided flyers which contained contact information for the study coordinator, and a URL for a web site that was set-up for students to self-enroll for test session slots. Upon enrollment (either by in-person or through the website) the participant was emailed a summary of the study procedures and a consent letter. They were asked to consider what they would like to use to create audible alerts and to bring their sounds or ideas about sounds to the study session.

No personal information was recorded other than contributed contact information.

At the beginning the participant was asked whether they had any formal musical training, since this was a factor that might influence performance: melodic alerts have been reported to be remembered better by musically trained subjects [3], [1].

The protocol comprises of two sub-studies:

1. Comparing performance of a subject with standard versus personalized audio alerts when discriminating within a set of six alerts, and then repeating this comparison with the same two sets of alerts in a second session after a gap of two weeks.

a. The subject, who is identified by a sequentially generated integer (a participant ID), is tested with both sets of sounds in a single session.

b. At the start of the first ("Baseline") session, the subject designates the sounds that are to be used for personalized alerts. These may be short phrases that the subject speaks into the microphone, music supplied by the subject as media files (e.g., WAV, MP3, MP4) from which samples of 2-10 seconds duration are made to the subject's specification, or a sound/music clip that is downloaded from an Internet resource such as YouTube and similarly cropped to 2-10 seconds duration. Each sound is designated by a number from 1 to 6, where 1 indicates the highest priority/ urgency, and 6 the lowest.

c. In half the participants (participants with an even-numbered participant ID) the set of standard alerts is employed first for testing, while in the other half, the set of personalized alerts is employed first.

d. Immediately prior to testing with a set of alerts, the participant has a "training run" of all six sounds played in sequence three times. The study coordinator reminds the participant that they may listen to each sound as long as they wish, and that the next sound will only be played when they touch the panel that is mapped to the sound. In the training session, if the subject does not select the correct alert, that alert will be repeated until selected correctly. The user sees a blank screen divided into six columns, numbered 1-6 for highest to lowest priority.

e. The test run for a set consists of 105 randomly chosen alerts. In 100 of these, single sounds are played. The time interval between alerts falls within a timing window that varies randomly between 100 milliseconds and 2 seconds after the user makes a response. At every twentieth (100/5=20) single-sound alert, two sounds are played randomly within the 2-second interval. For the simultaneous alerts, the subject is instructed to attempt to respond to the highest-priority alert first, and the program goes back into single-sound mode only after the user has made two responses.

f. The results recorded during the experiment (described in section 3.2.3) are stored within the application in a dedicated table space.

g. The purpose of the second ("Repeat") session is to measure retention in long-term memory, in order to test the hypothesis that alerts that have been personalized to a subject's preferences are recalled better than standard alerts. At the start of each set in the Repeat session, the experimenter sounds the alerts just once in sequence, at a 3-second gap between sounds, as a memory refresher. Such prompting, which primes the subject's mind for recall, is a legitimate intervention: it is too short to trigger another cycle of learning in the vast majority of subjects, who must rely instead on their long-term recall. Priming is necessary because the subjects have had no opportunity to rehearse their memory of the alerts after the Baseline session. The rest of the experiment is identical to the protocol as described for the baseline session.

2. The second study compares performance of a subject with standard versus personalized audio alerts when discriminating within a set of ten alerts. The experimental design is identical to that of the Baseline session of the first experiment, with the difference that the user interface is divided into two rows of five areas each (see Figure 2), and each of the ten icons indicates a specific type of critical care alert. The icons, in order from highest to lowest priority are:

a. Heart alert
b. Oxygen alert
c. Ventilator alert
d. Power failure alert
e. Perfusion alert
f. Drug/Medication alert
g. Temperature alert
h. General Alert, High Priority
i. General Alert, High Priority
j. Information Alert

There is no repeat session.

3. Categorization of Sounds used by Subjects for Personalized Alerts

Immediately after the baseline session concludes, the experimenter listens to the sounds designated by the subject for the personalized alerts, and categorizes them into four groups:

i. Musical tones, including segments of recorded music.
ii. Phrases recorded in the subject's own voice, or voice audio clips.
iii. Abstract sounds – e.g., sound effects from cartoons, simple chimes or rings, etc.
iv. Auditory Icons (“Earcons"): these are sounds that are auditory equivalents of the icons presented in the user interface, where the sound directly suggests the nature of the alert: Edworthy et al [23] refer to these as "auditory metaphors". They apply only to the 10-alert experiment. Examples are the rhythmic sound of heartbeats, bubbling sounds to indicate oxygen, etc. One subject selected a sample of Darth Vader's breathing from a "Star Wars" clip on YouTube. Many of the visual icons corresponding to an alert (e.g., the

temperature or medication alerts) do not have an intuitively obvious auditory icon.

3.4 Rationale of Experimental Design

3.4.1 Study 1: Six Alerts, Personalized vs. Standard, Repeated after 2 weeks

There are numerous variables that can affect performance in studies of recall, such as the sounds or music used for the alert, and the subject's general intelligence, musical ability, and memorization skills. Using a crossover (paired) design, with each subject as their own control, all factors other than the variable that we are testing – the choice of alert (the "intervention") – are controlled and eliminated in the comparison.

The choice of six alerts was made for two reasons:

- George Miller's highly cited 1956 paper, "The Magic Number seven, plus or minus two" [39] indicated that that the number of unrelated objects ("chunks") that an average human can hold in working memory is seven: later work indicated that the number of chunks depends on the category (i.e., span is seven for digits, six for letters, and around five for words.)
- Our preliminary experiments with standard alerts on volunteers prior to conducting the test indicated that six is a reasonable number in terms of cognitive burden, the memorization task being neither insuperable nor too easy.

It is our hypothesis that the participants' response to the personalized alerts will be faster and more accurate than the standard alerts. PASA records the response time to individual alerts is measured by total response time.

3.4.1.1 1b. Sub-analysis of Error rate with Simultaneous Alerts

As stated earlier in the description of the experimental protocol, in each session, after every 20th alert, the PASA software plays two alerts one after the other, and the subject is expected to make two responses in succession, indicating the higher-priority alert first and the lower-priority alert next. This is a simulation of a complex cognitive

task that occurs occasionally in real scenarios, where two alerts associated with a given patient may go off, and the healthcare provider is expected to attend to both (but attend to the more urgent one first). As stated earlier, one of the complaints against the ISO 60601-8 standard alerts was that the high-priority alerts are essentially the same as medium-priority alerts with the addition of two extra tones at the end. When both types of alerts fire near-simultaneously (and do so repeatedly), the subject is often likely to notice only one of the alerts (typically the higher-priority alert, because of the additional tones), and the medium-priority alert can be masked. One may therefore expect that the error rate with simultaneous alerts in general will be significantly higher than the overall error rate: it is worth studying whether personalization strategies, which typically result in subjects choosing sounds that are very distinct from each other, would reduce the error rate. One limitation of the current version of the software is that if the subject identifies both alerts correctly, but transposes the responses so that the lower-priority alert is responded to first, the alerts will both register correct as response order is not implicitly identified. This was done to allow correct alerts to register in the event that the participant responded to one alert before the second one sounded; alert events are still scored as individual events in a simultaneous alert.

When performing the repeat session after a two-week gap, the subject has had minimal time to rehearse the alert sounds, relying on long-term memory, and the difficulty in responding to both of the simultaneous alerts in correct order may be somewhat magnified. We need to test whether this is in fact the case.

3.4.2 Study 2: Ten Alerts with Icons, Personalized vs. Standard

The ten-alert experiment was intended to simulate a scenario that more closely and realistically approximates the role of healthcare providers in ICUs, who have to deal with as many as 12-16 different types of audio alerts. The "7± 2" threshold was deliberately exceeded, posing a challenge to most subjects' capabilities of short-term

memorization. The user interface was altered based on feedback from one member of the thesis committee during presentation of the thesis proposal and demonstration of an earlier version of PASA. The concern raised was that indicating target areas on the touch-screen interface with numbers only had a drawback: realistic control panels, e.g., as used in automobile dashboards or aircraft control panels use icons. Without icons, the subject has to make an explicit mental association between a specific area on the screen (i.e., a number) and a specific alert type. This results in an additional cognitive burden when the number of alerts increases, especially given the short training time in which the subject is expected to learn the alert sounds. Icons that actually represented the nature of the alert were therefore incorporated into the user interface, as stated earlier.

As discussed in the Results section, our data from the six-alert baseline experiment indicated that, on average, our subjects, while getting only four personalized alerts wrong in a session, were getting almost 27 standard audio alerts wrong. The potential for confusion and errors in dealing with many more types of alerts is far greater: many of these alerts sound so similar that, in actual clinical scenarios, they may only provide information about the general direction where the alert is originating: once at the bedside, the healthcare provider generally has to spend additional time making a more precise diagnosis.

The entire purpose of personalized alerts is to reduce cognitive burden on the user and greatly reduce (or even eliminate) training time, in the same way that, while manually deciphering a message that is being transmitted in Morse code requires specialized knowledge of a set of 36 symbols of dashes and dots, reading the equivalent plaintext (where technology has performed the signal translation) requires only basic literacy. Through the 10-alert experiment, we wished to determine whether, as the number of distinct alerts increases to the point where it poses a challenge to memory, subjects who have the freedom to personalize their alerts use personalization approaches that minimize memorization burden. An ancillary question to be explored is whether

these strategies differ from those used in the 6-alert situation, where the number of distinct alerts is small enough to be readily memorized by the average subject.

3.5 Statistical Analysis: Methodology and Justification

3.5.1 Prior to the Experiments: Sample Size Estimates

When we applied to the IRB for permission to recruit subjects, we performed a power analysis/sample size estimation using the freeware program PS from Vanderbilt University's Department of Statistics [40]. We estimated that if we assumed a difference between the control (standard) vs. test (personalized) interventions of 0.33 standard deviations, with an alpha error of 0.05 and a power of 80%, we would require a minimum of 73 subjects This number represents an upper bound: if the difference between the alert types were larger than this, the number of subjects required to demonstrate a difference between groups would be correspondingly fewer.

As discussed in the Results, however, our estimate of a 33% difference turned out to be highly pessimistic. On analyzing our data after 8 subjects, we found a ratio of error rates for standard versus personalized alerts approaching 6:1 on average, with personalized alerts providing consistently better performance in every subject tested. We were consequently able to reach high statistical significance ($p = 0.00185$ for errors and $p=0.00776$ for response time) with only 8 subjects, and we decided to stop after testing the 11 subjects we had already scheduled, and moved on to testing a 10-alert scenario.

3.5.2 Data Analysis after Completion of Experiments

The appropriate choice of analytical method for experiments such as ours depends on whether the distributions of the critical parameters – error rate and reaction times – would fit Gaussian (normal) distribution patterns when measured across subjects. After our experiments were completed, we employed the Anderson-Darling test of normality.

We found that the data for reaction times fit the assumption of normality. We therefore present reaction-time summary data using arithmetic mean, standard deviation and range. Levene's test for equality of variances showed that the variance of the reaction

time data for the 6-alert data did not differ significantly from that of the 10-alert data. When comparing six-alert data to 10-alert data for reaction time, we therefore employ the unpaired t test with assumption of equality of variances. We used a paired t-test for within-subject comparisons (e.g., personalized vs. standard alerts, baseline vs. repeat).

However, the error data significantly departed from a Gaussian distribution ($p < 0.005$ for both the 6-alert baseline and 10-alert experiments for personalized alerts). For example, with personalized alerts, while most subjects had seven or less errors in the six-alert baseline experiment, one subject had 33 errors. When we present summary data for errors, therefore, we present medians and ranges, rather than using means and standard deviations. For comparisons of error data, we use the Wilcoxon signed-rank test (the non-parametric analog of the paired t-test) for within-subject comparisons. To compare performance across subjects (6 alerts versus 10 alerts), we employ the Mann-Whitney test (the non-parametric analog of the unpaired t-test) for error data.

We used Minitab version 16 (Minitab Inc., State College, PA) for all the data analyses.

CHAPTER 4 RESULTS

4.1 Standard vs. Personalized Alerts, 6 Alerts: Baseline Study

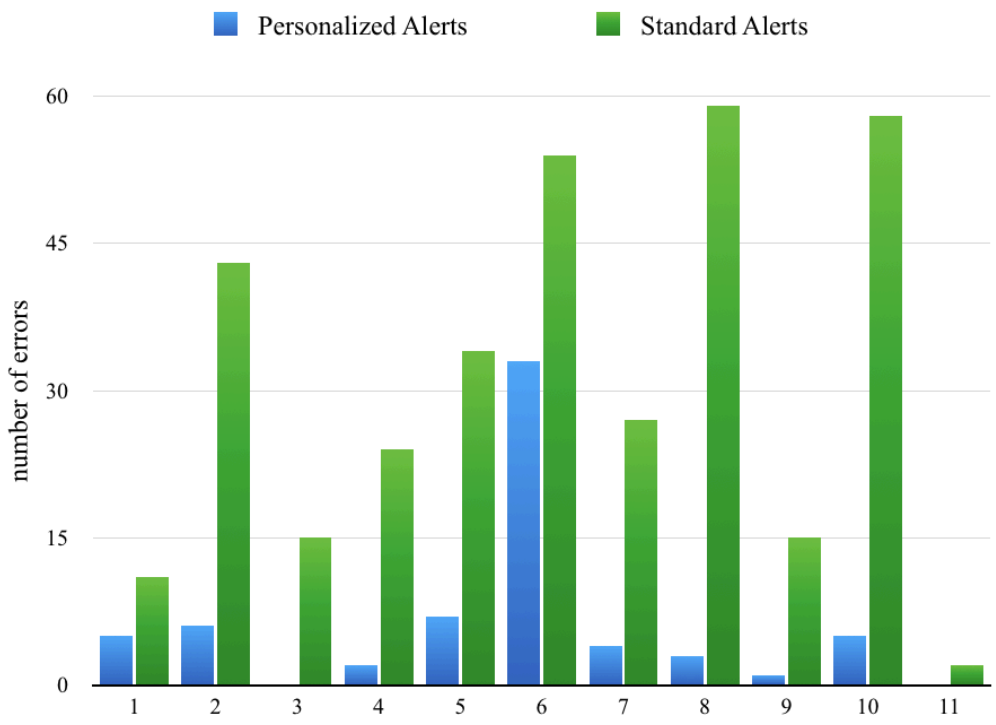

Figure 14. Errors: 6-alerts

4.1.1 Errors for session 1: training and test

Table 4. Errors across session 1: training and test

Measure	Personalized Alerts	Standard Alerts
Median number of errors and range	4 (0-33)	27 (2-59)
P value for differences by Wilcoxon signed-rank test: 0.004		

See table 4 and figure 14. While performance in terms of errors varied greatly across subjects, the error rate with personalized alerts was consistently less than for standard alerts in all 11 subjects.

4.1.2 Reaction Time

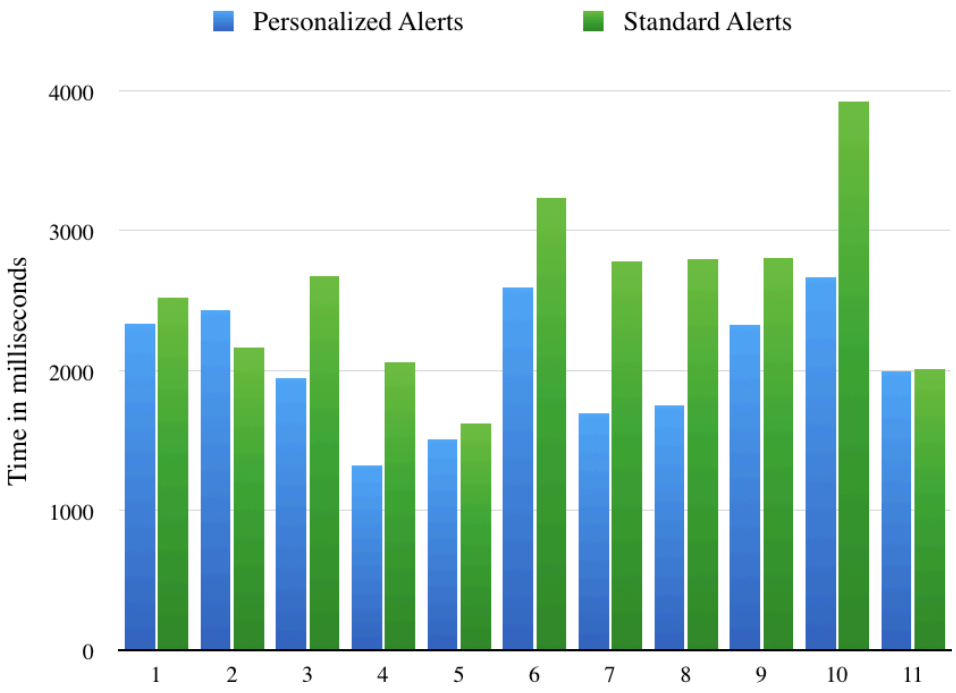

Figure 15. Reaction time: 6-alerts

Table 5. Reaction times: 6-alerts

Measure	Personalized Alerts	Standard Alerts
Mean ± Standard Deviation of average reaction time (in milliseconds), with range across subjects	2051 ± 451 (range 1317-2668)	2559 ± 638 (range 1623-3921)
P value for differences by paired t-test: 0.004		

Table 6. Errors for simultaneous alerts: 6-alerts

Measure	Personalized Alerts	Standard Alerts
Median number of errors and range	0 (0-1)	5 (0-10)
P value for differences by Wilcoxon's Signed-Rank Test: 0.006		

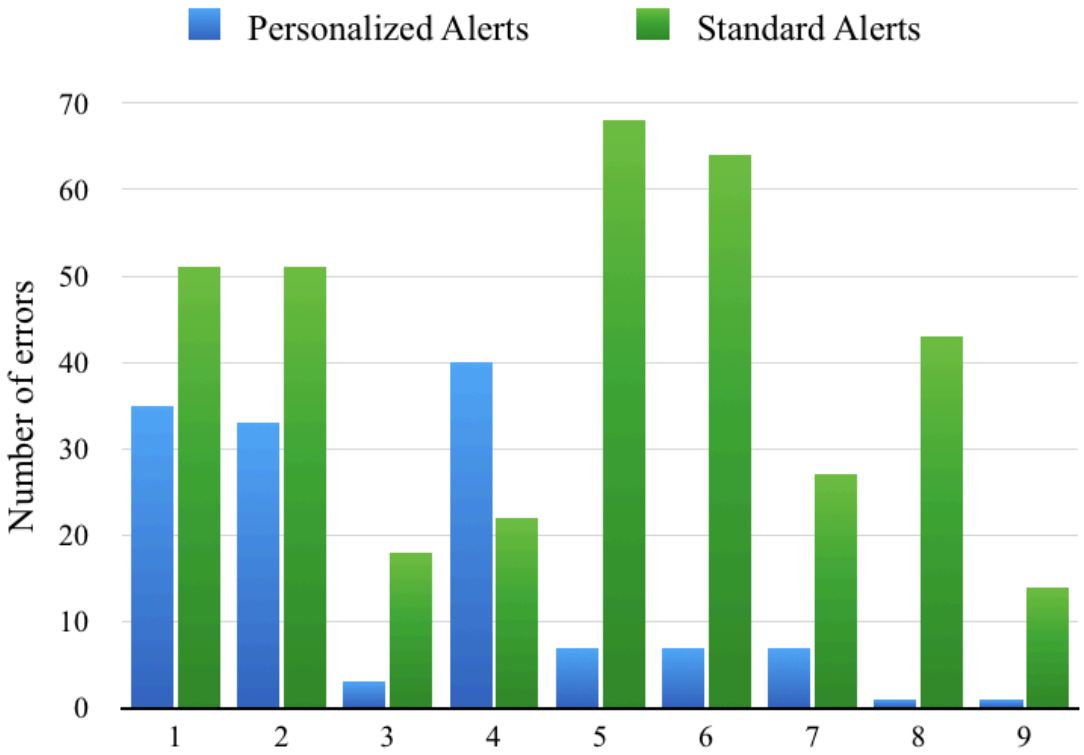

Figure 16. Repeat study: 6-alerts

See table 5 and figure 15. 10 of 11 subjects had shorter average reaction time with personalized alerts than with standard alerts. In the remaining subject, the numbers were 2433 vs. 2163 milliseconds for personalized vs. standard alerts. This subject's errors, however, were 6 for personalized vs. 43 for standard. Reaction time is less important than accuracy in this context, and in most other contexts as well: one of Jon Bentley's tongue-

in-cheek dictums with respect to software development and algorithm performance (stated in his classic collection of columns, "Programming Pearls" [41]), goes: "If a program or human is not required to compute a correct answer, one can achieve almost infinite performance gain by simply generating a random answer".

4.1.3 Errors for Simultaneous Alerts: 6-alerts

See table 6. With personalized alerts, 9 out of 11 subjects had no errors at all: the remaining two had 1 error each. With standard alerts, on average the subjects got half the responses wrong.

4.2 Standard vs. Personalized Alerts– 6 Alerts: Repeat Study

Two of the 11 subjects who took the baseline study were unable to perform the repeat study. We therefore illustrate the results only for the remaining nine. As shown from the numbers below, the drop-outs have not influenced the outcome of the experiment materially: the differences between performance for personalized versus standard alerts continues to be demonstrated with only slightly larger p values.

4.2.1 Errors across the session 2: repeat without training

See table 7 and figure 16. While performance in terms of errors varied greatly across subjects, the error rate with personalized alerts was less than for standard alerts in 8 of 9 subjects.

Table 7. Errors across session 2: repeat without training

Measure	Personalized Alerts	Standard Alerts
Median number of errors and range across subjects	7 (1-40)	43 (14-68)
P value for differences by Wilcoxon's Signed-Rank Test: 0.038		

Table 8. Reaction time across session 2: repeat without training

Measure	Personalized Alerts	Standard Alerts
Mean and Standard Deviation of average reaction time (in milliseconds), with range across subjects	2776 ± 446 (range 1805-3259)	3377 ± 389 (range 2778-3886)
P value for differences by paired t-test: 0.0018		

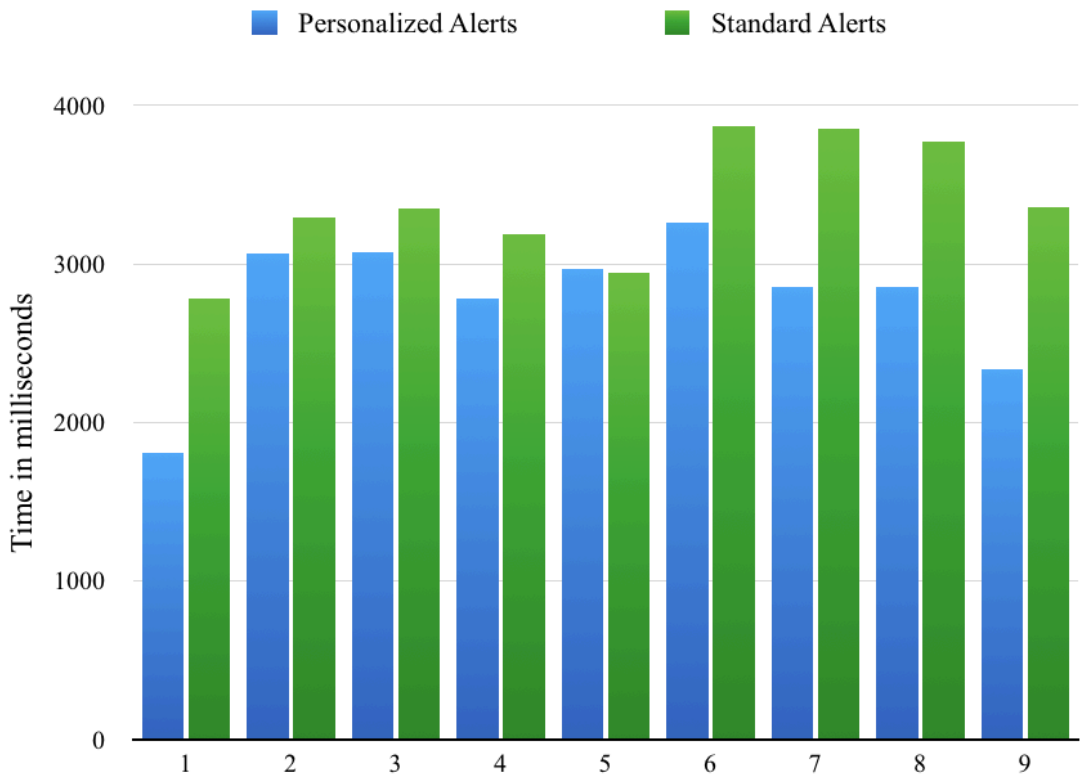

Figure 17. Reaction time for repeats: 6-alerts

Table 9. Errors for simultaneous alerts

Measure	Personalized Alerts	Standard Alerts
Median number of errors and range across subjects	0 (0-1)	5 (5-6)
P value for differences by Wilcoxon's Signed-Rank Test: 0.009		

4.2.2 Reaction Time

See table 8 and figure 17. 8 of 9 subjects had shorter reaction time with personalized alerts than with standard; the remaining subject's numbers were 2965 ms for personalized vs. 2941 for standard, with corresponding error rates of 7 versus 64. Again, as for the Baseline study, accuracy is more important than reaction time.

4.2.3 Errors for Simultaneous Alerts

See table 9. With personalized alerts, 9 out of 11 subjects had no errors at all: the remaining two had 1 error each. With standard alerts, on average the subjects got half the responses wrong.

4.3 Musical Training vs. Error Rate for Standard and Personalized Alerts: 6-alert experiment

Of our 11 subjects, 8 stated that they had some degree of musical training: this proportion is probably fairly typical for the state of Iowa, where most subjects have had some exposure to band training in middle and high school. While the lowest error rate overall (zero errors with personalized alerts, only 2 errors with standard alerts) was with a subject who had majored in classical music, we found no statistically significant association between musical training and errors for either the standard alerts or the personalized alerts when these were analyzed separately, which corresponds to similar

findings in past research with alerts that were not melodic [19], but does not correspond to results reported in the literature for melodic alerts.

While our results do not rule out such an association, which has been reported in the literature for alerts meeting the 60601-1-8 specification, as stated previously, this may be a consequence of the small number of subjects tested, with only 3 non-musically trained subjects. In any case, this association was not a r hypothesis that we intended to spend resources testing by recruiting additional subjects – the entire objective of personalization of alerts is that lack of musical training should not be a barrier to the ability to function effectively in ICU settings. The best that we can state, given our data, is that any effect that musical training may have on performance is dwarfed by the effect of personalization.

4.4 Contrast of Baseline Experiment with Repeat Experiment

See table 10. These results are only reported for the 9 subjects who performed both the Baseline and the Repeat Study. We present the comparisons in a single table (see table 10), with relatively minimal comment because the results are not particularly surprising: subjects perform more slowly /cautiously during the repeat session than during the baseline session because they are required to call on their long-term memory. The differences between the errors, however, do not reach statistical significance even though the worst-case performance with the Repeat study is poorer than with the Baseline study: a few subjects actually show modest improvement or no change in the Repeat study. Testing degradation of performance accuracy, if any, after an attempt at recall was not a hypothesis of this study, in any case, and we only conclude that the effect of a temporal delay on performance is too small to be picked up with the present sample size. Also, the subjects received an initial prompt by hearing the alerts in sequence just once immediately prior to the testing, and the act of prompting may have prevented significant accuracy degradation in several subjects.

Table 10. Baseline with repeat: 6-alert test

Experimental parameter	**Summary Statistics**	**Baseline**	**Repeat**	**P value for comparison**	**Comments**
Personalized alerts: Errors, overall	Median number of errors and range across subjects	4 (0-33)	7(1-40)	P = 0.15 (not significant)	2 subjects had slightly better performance on repeat, 2 had no change
Personalized alerts: Average Reaction time	Mean, standard deviation and range across subjects	1988 ± 450 (1317-2591)	2776 ± 446 (1805-3259)	P = 0.008 by paired t-test	
Personalized alerts: Errors, simultaneous alerts	Median number of errors and range across subjects	0 (0-1)	0 (0-1)	Differences not significant by Wilcoxon signed rank	
Standard alerts: Errors, overall	Median number of errors and range across subjects	27 (11-59)	43 (14-68)	P = 0.234 (not significant) by Wilcoxon's signed rank test	3 subjects had slight improvement, 1 had no change
Standard alerts: Average Reaction time	Mean, standard deviation and range across subjects	2518 ± 489 (1623-3238)	3377 ± 389 (2778-3866)	$P = 7.43 \times 10^{-5}$ by paired t-test	
Standard alerts: Errors, simultaneous alerts	Median number of errors and range across subjects	5 (5-10)	5 (5-6)	Differences not significant by Wilcoxon signed rank	

4.5 Standard vs. Personalized Alerts– 10 Alert Experiment

13 subjects participated in this experiment. As stated earlier, apart from increasing the number of distinct alerts, icons indicating the nature of each alert are placed in corresponding sections of the touch-screen interface.

4.5.1 Errors across the session: training and test

Table 11. Errors across session: training and test

Measure	Personalized Alerts	Standard Alerts
Median number of errors and range	1 (0-13)	55 (15-95)
P value for differences by Wilcoxon signed-rank test: 0.002		

Differences between personalized alerts and standard alerts were even more significant for the six-alert sessions: across subjects, the worst performance with personalized alerts was better than the best performance with standard alerts. See table 11.

4.5.2 Reaction Time

Table 12. Reaction time

Measure	Personalized Alerts	Standard Alerts
Mean ± Standard Deviation of average reaction time (in milliseconds), with range across subjects	2971 ± 364 (range 2322-3761)	4356 ± 639 (range 3263-5569)
P value for differences by paired t-test: 2.2×10^{-12}		

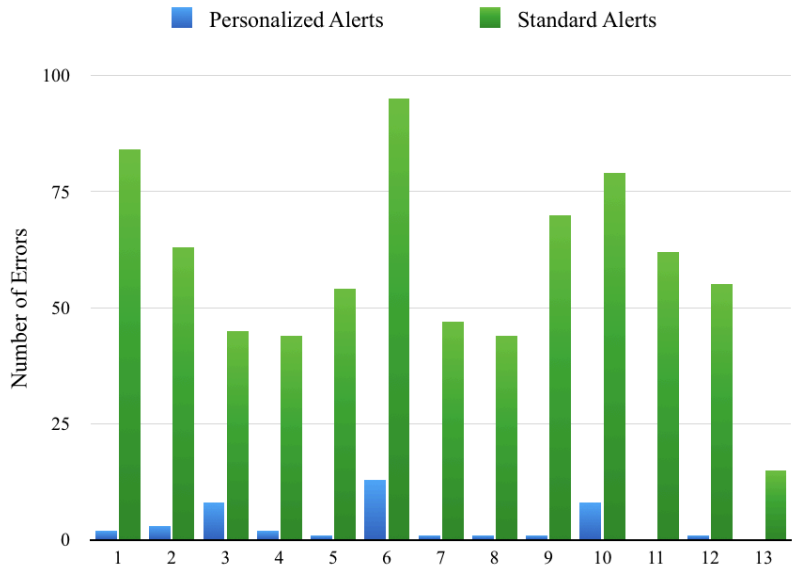

Figure 19. Errors: 10-alerts

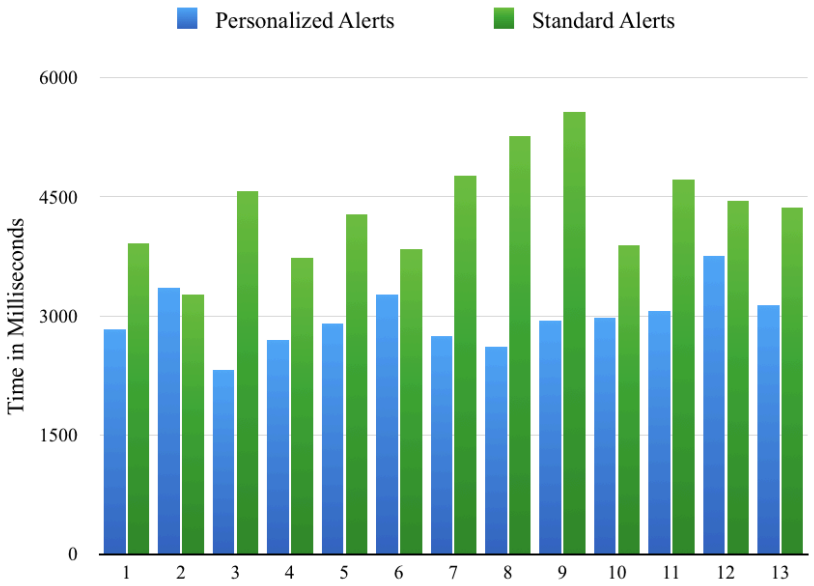

Figure 18. Reaction time: 10-alerts

12 of 13 subjects had shorter average reaction time with personalized alerts than with standard alerts. In the remaining subject, the numbers were 3352 vs. 3264

milliseconds for personalized vs. standard alerts, with corresponding errors of 3 vs. 63. See table 12 and figure 18.

4.5.3 Errors for Simultaneous Alerts

Table 13. Errors for simultaneous alerts

Measure	Personalized Alerts	Standard Alerts
Median number of errors and range	0 (0-2)	4 (0-10)
P value for differences by Wilcoxon's Signed-Rank Test: 0.002		

With personalized alerts, 8 out of 13 subjects had no errors at all, 4 had one error and 1 had 2 errors. With standard alerts, on average the subjects got half the responses wrong. See table 13 and figure 19.

4.6 Standard Alert Recognition Rates

As stated earlier, a common complaint about the 60601-1-8 standard is that the 5-note high-priority alerts are very similar to the 3-note medium-priority alerts. The low-priority alerts, which have only 2 notes, may be easier to remember (though they may also be confused with each other, especially when there are many of them). From our data, we wished to test the hypothesis that, in the pooled subject data for a given session, the errors across individual alerts are non-uniform.

In the tables presented below (see tables 14 and 15), we present the percentage of alerts that were incorrectly identified, by alert. (The raw numbers of errors and correct responses have not been presented to simplify interpretation: due to the random nature of alert generation, the numbers of individual alerts presented to the subjects are unequal across alerts).

Table 14. 6-alert data: pooled baseline and repeat data

Target Alert	1	2	3	4	5	6
Error Percent	29.1	29	42.8	41.7	47.6	21.1

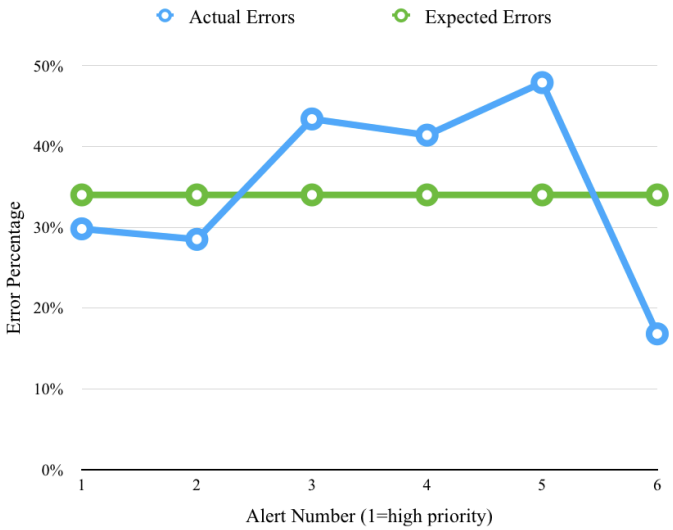

Figure 20. Observed vs. expected error percentages (6-alerts)

Table 15. 10-alert data

Target Alert	1	2	3	4	5	6	7	8	9	10
Error Percent	58	63	66	67	63	77	35	37	52	45

In both cases, the distributions were non-uniform, as shown by figures 20 and 21. We tested for uniformity by using the chi-squared statistic: the "observed" values were the raw numbers of errors and correct responses, while the "expected values", assuming a uniform distribution of error across alerts, were computed by the simple formula (total errors (or correct responses)/ number of distinct alerts).

The respective p values for the 6 alert set data and the 10 alert data were 2.7562 x 10^{-17} and 9.979 x 10^{-12} respectively. The error rate for both peaks for the medium priority alerts. This may be due to the fact that the medium priority alerts are easily confused with the high priority alerts.

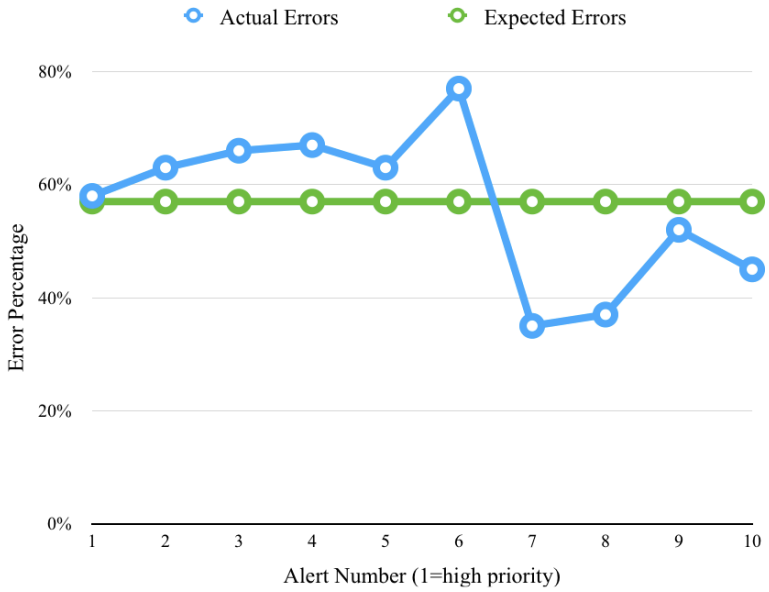

Figure 21. Observed vs. expected error percentages (10-alerts)

4.7 Comparison of Error Rate and Reaction Times for 6 Alerts vs. 10 Alerts

See table 16. We present the summary statistics and comparisons in a single table analogous to Table 8, noting that these numbers reflect different sets of subjects. We have omitted the sub-comparisons for simultaneous alerts, which were not statistically significant. The degradation in performance with standard alerts as the number of distinct alerts is increased, as expected. Of interest, however, are the numbers for personalized alerts. While the reaction time is slower (as expected), the errors do not worsen: in fact,

Table 16. Error rate and reaction time for 6-alerts vs. 10-alerts

Experimental parameter	Summary Statistics	6 Alerts baseline	10 Alerts	P value for comparison
Personalized alerts: Errors, overall	Median number of errors and range across subjects	4 (0-33)	1 (0-13)	P not significant by Mann-Whitney test
Personalized alerts: Average Reaction time	Mean , standard deviation and range across subjects	2051 ± 451 (1317-2668)	2971 ± 364 (2322-3761)	P = 0.008 by unpaired t-test
Standard alerts: Errors, overall	Median number of errors and range across subjects	27 (2-59)	55 (15-95)	P = 0.0071 by Mann-Whitney test
Standard alerts: Average Reaction time	Mean , standard deviation and range across subjects	2559 ± 638 (1623-3921)	4356 ± 640 (range 3263-5569)	$P=1.3x10^{-16}$ by unpaired t-test

while the differences are not statistically significant, the median number of errors with 10 alerts (1 error) is actually *less* than for six alerts (4 errors). Though this result may seem paradoxical, it is actually not, as discussed in the next subsection. The use of icons in the user interface that suggest specific alerts allows subjects to choose personalization strategies that maximize the likelihood of not making a mistake.

4.8 Personalization Strategies: Comparison

As stated earlier, we have categorized the personalized-alert sounds that subjects chose into four categories: voice (spoken words or phrases), musical segments, abstract sounds and earcons/auditory icons. Table 17 shows the distribution of categories in each study.

Table 17. Personal alert categories

Study	Voice	Music	Abstract	Earcons	Total
6 Alerts	29	25	12	0	66
10 alerts (with icons)	78	20	16	16	130
Totals	107	45	28	16	196

P value by Chi-square test = 0.00011.

The p value above shows that the strategies employed by subjects differ based on the experimental scenario. For the 6-alert study, participants more quickly selected sounds for the high, medium, and low priority alerts than the 10-alert study participants, who selected sounds that would correspond to the icons representing physiological conditions.

For the 10-alert study, participants were more careful and selective when creating sounds, often referring to the iPad interface to make sure that they understood the functions that the individual icons represented. When icons are presented in the user interface, the user preference shifts towards voice – short phrases in the subject's voice

that directly name the alert ("heart", "oxygen", "perfusion", "meds"), with some subjects choosing sounds – heartbeats, heavy breathing, the sound of a fan – that directly suggest the alert.

Others chose music exclusively: two subjects chose classical music segments, while another chose theme song clips from Disney movies. Horns and sirens were commonly selected for the highest priority alerts; one participant wanted a very specific siren that was used in a scene from the movie “Despicable Me 2”, another wanted the opening to Beethoven’s 5th symphony. For the oxygen alert in the 10-alert study, bubbles were often requested, and one participant wanted the sound of Darth Vader breathing. For perfusion, dripping water was commonly requested.

For low priority alerts, softer music or sounds were the norm, with “low priority” or “information only” occasionally vocalized.

With the exception of three participants, everyone chose to record themselves saying “drugs” for the drug alert. Of the three exceptions, one was the person who exclusively used Disney theme songs, and the other exceptions were the two who chose all classical music.

Participants who had spent time in a clinical setting were more apt to use voice recordings; for example, “code” was assigned to the cardiac alert, and “check vent” was used for the ventilation alert.

Many of the participants came into their sessions with a clear idea of what they wanted for some, if not all, of their sounds; many had lists prepared, and one participant came with a USB drive upon which they had already placed some sound files that they wanted to use.

For the 10-alert study, voice alerts were used the most (78 out of 130 alerts), which might imply that as the number of alerts to be recognized increases, the reliance on phrases will go up. For example, in an actual clinical-care setting, a healthcare provider who is managing several patients, given the ability to customize alerts for individual

sensor devices from a single console, might choose a phrase such as "Patient Jane Smith, oxygen": in such cases, the auditory alert conveys maximal information that requires no additional cognitive effort to process beyond that required to understand one's own speech.

4.9 Discussion

4.9.1 Key observations

It seems clear that personalized alerts provide an opportunity for improving information management at the point of patient care. In our study, personalized alerts were much more accurately retained and recalled in both the 6-alert and 10-alert tests Response times, while not showing quite the ratio of difference exhibited in recollection and retention, were still better for personalized alerts than standard alerts. These results, while expected, were surprisingly decisive with respect to the potential that personalized alerts may have in health care. Our study sheds light on a new modality for alert design that merits continued study- not only with existing patient monitoring information technology (IT) artifacts, but especially new IT artifacts that are being explored for usage in point of patient care; smartphones, tablets, and wearable computers.

At the end of their study sessions, all of our participants expressed a better understanding of the complexity and challenges that audio alerts can impose in a critical care setting, and while they all “got it”, the participants who had actually spent time working in critical care environments were especially adamant expressing appreciation for our research, further adding credibility to the value of continued research of this type.

4.9.2 Additional Study Observations

As a tool for constructing audio alerts PASA worked quite well. All requests for sound sources were accommodated, and sound editing was intuitive enough to allow study participants to understand what was taking place and simple enough to inspire participants to feel comfortable asking questions and not settle for anything less than their “ideal” alerts.

When, after looking at the results from the 6-alert test, we decided to move to a 10-alert test, the study designer allowed us to quickly make the shift, and it took less than an hour to modify the PASA framework to change a single row format to double rows with icons for the 10-alert iPad interface.

The sound strategies employed by the participants, may form the focus of future experiments. While the 6-alert and 10-alert participants often shared commonalities between types of sounds that were used for certain alerts such as highest priority and lowest priority, the 10-alert participants were more deliberate in their sound creation. They often referred to the icons on the iPad and asked enough questions to be sure of the physiological functions they represented. Also, while the 6-alert participants usually created their alerts in priority sequence, the 10-alert participants did not always create their alerts in sequence; they often started with alerts that they understood best, or those that immediately stood out in their minds as a candidate for a match with a specific icon.

4.9.3 Future experiments with Personalized Alerts designed with PASA.

A future study protocol might involve making the icons an independent variable. The students might be asked to create alerts that meet one of the physiological categories and then choose from a selection of icons for each category that they perceive matches best. In such an experiment, the control group might assign alerts to predetermined icons, while the test group would get to choose their own icons. It would be interesting to see if the personalization of icons correlates to an increase in response rate. Such personalization might also help our ability to quickly make an informed decision during a critical care alarm event, and certainly would not be unreasonable if future EHR systems accommodated information delivery that was situationally contextualized.

A key influence behind the design of PASA was to extend personalized alert study protocols into health care simulation protocols. PASA can control several iPad interfaces simultaneously, and those interfaces can be highly accurate touchscreen representations of actual medical equipment. PASA could serve as a controller for a study

protocol where one of the study groups could be shown biometric information and hear alerts through a device such as Google Glass. iPads could be placed where actual medical devices exist and, by simulating the functionality of the device with fully functional controls such as buttons sliders, etc., PASA could be used to generate alarm scenarios and record reaction and response times, as well as action taken by the participant. This would allow various strategies for information delivery at the point of patient care to be tested in an environment that approximates actual clinical care settings. In 2012, Bennett and McNeer built a tool for designing audible alerts and integrating the alert annunciation function with actual medical devices. Their tool required expertise in C++, Matlab, and the ability to interface directly into the medical devices. It also had 11 audio drop-outs when the number of alarms exceeded 16; problems that do not exist within the PASA framework. However, their findings indicate that a tool that could rapidly stage multiple audible alarm sets from the laboratory to a simulation environment for the purpose of evaluating novel alarm designs could produce valuable findings for medical audible alarm standardization; an important consideration since IEC stipulates that all new alarm designs should be subjected to clinical or simulated clinical usability testing [42].

A more simple use of PASA would be to make existing simulation more realistic. A student holding an iPad and facing a mannequin could have a number of buttons that are labeled with questions that could be asked, such as “how are you feeling”, “what hurts”, or, “what medications are you on”, etc. A video of an actor could then appear on the iPad screen that exhibits the symptoms and answers the questions. This would add an aspect of realism to a mannequin based simulation, maximize the utility of human actors performances, and also serve as another means to quantitatively measure performance [43].

Because the iPad interface is controlled entirely by PASA, many of issues related to acquiring data from a stand-alone iPad application are eliminated, e.g, synchronizing the tablet, merging the data, assigning data to the proper study participant, etc. Since

PASA serves as the controller for the simulation and the date repository for measurement metrics for an experiment, instructors and students could review response data to a simulation immediately upon completion of the simulation. Building this capability is more labor intensive and requires much broader programming skills in web applications or stand-alone applications that run natively on the iPad.

4.9.4 PASA in other contexts

PASA could also have utility in other areas beside personalized alerts. With minor modification, PASA could be used as a tool for teaching audio-to-visual mapping skills; something that would benefit autistic children [44]–[49] and elderly patients suffering from disabilities such as Alzheimer's, stroke, or age-related dementia [50]–[53].

It would also be interesting to use PASA to build studies that test various theories for learning alert sounds. While we trained our participants by cycling through alerts that played in priority order a number of times, different training procedures could also be tested for retention characteristics. One such candidate would be the training sequence suggested by Patterson [13] where single sounds are played with additional sounds being aggregated to the sequence whenever recognition rates rise to a certain level. This would allow individualized training routines to be developed for a population, as one training methodology might not be optimal for everyone.

REFERENCES

[1] P. M. Sanderson, A. Wee, and P. Lacherez, “Learnability and discriminability of melodic medical equipment alarms*,” *Anaesthesia*, vol. 61, no. 2, pp. 142–147, 2006.

[2] A. N. Wee and P. M. Sanderson, “Are Melodic Medical Equipment Alarms Easily Learned?,” *Anesth Analg*, vol. 106, no. 2, pp. 501–508, Feb. 2008.

[3] P. Lacherez, E. L. Seah, and P. Sanderson, “Overlapping melodic alarms are almost indiscriminable,” *Hum Factors*, vol. 49, no. 4, pp. 637–645, Aug. 2007.

[4] D. J. H. Kerr and B. Hayes, “An ‘Alarming’ situation in the intensive therapy unit,” *Intensive Care Med*, vol. 9, no. 3, pp. 103–104, May 1983.

[5] S. Siebig, S. Kuhls, M. Imhoff, J. Langgartner, M. Reng, J. Schölmerich, U. Gather, and C. E. Wrede, “Collection of annotated data in a clinical validation study for alarm algorithms in intensive care—a methodologic framework,” *Journal of Critical Care*, vol. 25, no. 1, pp. 128–135, Mar. 2010.

[6] S. Siebig, S. Kuhls, M. Imhoff, U. Gather, J. Schölmerich, and C. E. Wrede, “Intensive care unit alarms--how many do we need?,” *Crit. Care Med.*, vol. 38, no. 2, pp. 451–456, Feb. 2010.

[7] “Too much noise from hospital alarms poses risk for patients,” *The Washington Post*. [Online]. Available: http://www.washingtonpost.com/sf/feature/wp/2013/07/07/too-much-noise-from-hospital-alarms-poses-risk-for-patients/. [Accessed: 08-Jul-2013].

[8] “State reports detail 11 patient deaths linked to alarm fatigue in Massachusetts,” *The Boston Globe*. [Online]. Available: http://www.bostonglobe.com/metro/2011/12/29/state-reports-detail-patient-deaths-linked-alarm-fatigue-massachusetts/C8y3itRvd4WiGnR40sVHxN/story.html.

[9] “Medical device alarm safety in hositals,” *Joint Commission Sentinal Event Alert*, no. 50, Apr. 2013.

[10] “Top 10 Health Technology Hazards for 2013.,” *ECRI Institute*, vol. 41, no. 11, Nov. 2012.

[11] R. D. Patterson, G. Britain, and C. A. Authority, *Guidelines for auditory warning systems on civil aircraft*. Civil Aviation Authority, 1982.

[12] R. D. Patterson, “Auditory warning systems for high-workload environments,” *Ergonomics International*, vol. 85, pp. 163–165, 1985.

[13] R. Patterson, J. Edworthy, M. Shailer, M. Lower, and P. Wheeler, “Alarm sounds for medical equipment in intensive care areas and operating theatres,” *Institute of Sound and Vibration Research, no. AC598*, 1986.

[14] “CSA-ISO 9703.2-97, ”Anaesthesia and respiratory care alarm signals - Part 2: Auditory alarm signals", Can Standards Assoc., 1997.

[15] IEC. International Standard IEC 60601 (2005): "Medical Electrical Equipment – Part 1–8: General Requirements for Safety – Collateral Standard. General Requirements, Tests and Guidance for Alarm Systems in Medical Electrical Equipment and Medical Electrical Systems." Geneva: International Electro-technical Commission, 2005.

[16] S. Williams and P. C. W. Beatty, “Measuring the performance of audible alarms for anaesthesia,” *Physiol Meas*, vol. 26, no. 4, pp. 571–581, Aug. 2005.

[17] F. E. Block, J. D. Rouse, M. Hakala, and C. L. Thompson, “A proposed new set of alarm sounds which satisfy standards and rationale to encode source information,” *Journal of clinical monitoring and computing*, vol. 16, no. 7, pp. 541–546, 2000.

[18] P. Sanderson, A. Wee, E. Seah, and P. Lacherez, “Auditory alarms, medical standards, and urgency,” *vascular (CV)*, vol. 4, pp. E4–G4, 2006.

[19] J. Allan, P. Milligan, and J. Cuthill, “‘Name that tune’. Alarms in intensive care-how good are we at recognising them?,” in *ANAESTHESIA*, 2011, vol. 66, pp. 412–412.

[20] F. E. Block, “‘For if the trumpet give an uncertain sound, who shall prepare himself to the battle?’ (I Corinthians 14:8, KJV),” *Anesth Analg*, vol. 106, no. 2, pp. 357–359, Feb. 2008.

[21] J. Edworthy and E. Hellier, “Alarms and human behaviour: implications for medical alarms,” *British Journal of Anaesthesia*, vol. 97, no. 1, pp. 12–17, 2006.

[22] J. Edworthy, E. Hellier, K. Titchener, A. Naweed, and R. Roels, “Heterogeneity in auditory alarm sets makes them easier to learn,” *International Journal of Industrial Ergonomics*, vol. 41, no. 2, pp. 136–146, Mar. 2011.

[23] J. Edworthy, “Designing effective alarm sounds,” *Biomedical Instrumentation & Technology*, vol. 45, no. 4, pp. 290–294, 2011.

[24] J. Edworthy, R. Page, A. Hibbard, S. Kyle, P. Ratnage, and S. Claydon, “Learning three sets of alarms for the same medical functions: A perspective on the difficulty of learning alarms specified in an international standard,” *Applied Ergonomics*, 2013.

[25] C. E. Osgood, “The nature and measurement of meaning,” *Psychological Bulletin*, vol. 49, no. 3, pp. 197–237, 1952.

[26] H. Von Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory or Music. ; Originally published as Die Lehrevon den Tonempfindungen als PhysiologischeGrundlagefur die Theorieder Musik (Braunschweig:F. Vieweg und Sohn, 1863)*. Longmans, Green, 1912.

[27] G. Von Bismarck, “Sharpness as an attribute of the timbre of steady sounds,” *Acta Acustica united with Acustica*, vol. 30, no. 3, pp. 159–172, 1974.

[28] R. L. Pratt and P. E. Doak, “A subjective rating scale for timbre,” *Journal of Sound and Vibration*, vol. 45, no. 3, pp. 317–328, 1976.

[29] E. C. Haas and J. G. Casali, “Perceived urgency of and response time to multi-tone and frequency-modulated warning signals in broadband noise,” *Ergonomics*, vol. 38, no. 11, pp. 2313–2326, 1995.

[30] E. C. Haas and J. Edworthy, “Designing urgency into auditory warnings using pitch, speed and loudness,” *Computing & Control Engineering Journal*, vol. 7, no. 4, pp. 193–198, 1996.

[31] C. Gonzalez, B. A. Lewis, and C. L. Baldwin, “Revisiting pulse rate, frequency and perceived urgency: have relationships changed and why?,” Jun-2012. .

[32] J. L. Burt, D. S. Bartolome, D. W. Burdette, and J. R. Comstock, “A psychophysiological evaluation of the perceived urgency of auditory warning signals,” *Ergonomics*, vol. 38, no. 11, pp. 2327–2340, 1995.

[33] N. Stanton and J. Edworthy, “Auditory affordances in the intensive treatment unit,” *Applied Ergonomics*, vol. 29, no. 5, pp. 389–394, Oct. 1998.

[34] S. Krug, *Don’t make me think!: a common sense approach to Web usability*. Pearson Education India, 2000.

[35] A. Testa, D. Gallo, and R. Langella, “On the Processing of harmonics and interharmonics: using Hanning window in standard framework,” *IEEE Transactions on Power Delivery*, vol. 19, no. 1, pp. 28–34, Jan. 2004.

[36] B. J. Fogg and H. Tseng, “The Elements of Computer Credibility,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 1999, pp. 80–87.

[37] S. Lee and S. Zhai, “The performance of touch screen soft buttons,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 309–318.

[38] “Cycling ’74. URL http://cycling74.com.” .

[39] G. A. Miller, “The magical number seven, plus or minus two: some limits on our capacity for processing information.,” *Psychological review*, vol. 63, no. 2, p. 81, 1956.

[40] W. D. Dupont and W. D. Plummer Jr., “Power and sample size calculations: A review and computer program,” *Controlled Clinical Trials*, vol. 11, no. 2, pp. 116–128, Apr. 1990.

[41] J. L. Bentley, *Programming pearls*. Addison-Wesley Professional, 2000.

[42] C. L. Bennett and R. R. McNeer, “PT-SAFE: A Software Tool for Development and Annunciation of Medical Audible Alarms,” *Anesthesia & Analgesia*, vol. 114, no. 3, pp. 576–583, Mar. 2012.

[43] “Emotions bring mannequin to ‘life,’” *BBC News*. [Online]. Available: http://www.bbc.com/news/technology-27492153. [Accessed: 24-May-2014].

[44] T. A. Cardon, “Teaching caregivers to implement video modeling imitation training via iPad for their children with autism,” *Research in Autism Spectrum Disorders*, vol. 6, no. 4, pp. 1389–1400, 2012.

[45] J. E. Hart and K. J. Whalon, “Using video self-modeling via iPads to increase academic responding of an adolescent with autism spectrum disorder and intellectual disability,” *Education and Training in Autism and Developmental Disabilities*, vol. 47, no. 4, p. 438, 2012.

[46] J. P. Hourcade, S. R. Williams, E. A. Miller, K. E. Huebner, and L. J. Liang, “Evaluation of tablet apps to encourage social interaction in children with autism spectrum disorders,” in *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, 2013, pp. 3197–3206.

[47] L. Neely, M. Rispoli, S. Camargo, H. Davis, and M. Boles, “The effect of instructional use of an iPad on challenging behavior and academic engagement for two students with autism,” *Research in Autism Spectrum Disorders*, vol. 7, no. 4, pp. 509–516, 2013.

[48] G. Trott, “iPads and Autism: The Impact of iPads on Social and Academic Development in Autism,” 2013.

[49] P. O’Malley, M. E. B. Lewis, C. Donehower, and D. Stone, “Effectiveness of Using iPads to Increase Academic Task Completion by Students with Autism,” *Universal Journal of Educational Research*, vol. 2, no. 1, pp. 90–97, 2014.

[50] J. Coppola PhD, M. A. Kowtko, C. Yamagata, and S. Joyce, “Applying Mobile Application Development to Help Dementia and Alzheimer Patients,” 2013.

[51] J. Coppola PhD, M. A. Kowtko, C. Yamagata, and S. Joyce, “Mobile app development and usability research to help dementia and Alzheimer patients,” 2013.

[52] C. Yamagata, J. F. Coppola, M. Kowtko, and S. Joyce, “Mobile app development and usability research to help dementia and Alzheimer patients,” in *Systems, Applications and Technology Conference (LISAT), 2013 IEEE Long Island*, 2013, pp. 1–6.

[53] J. F. Coppola, M. A. Kowtko, C. Yamagata, and S. Joyce, “Applying Mobile Application Development to Help Dementia and Alzheimer Patients.”