2007

Privacy Protection for Life-log System

Jayashri S. Chaudhari
University of Kentucky, jayashrirane@yahoo.com

ABSTRACT OF THESIS

Privacy Protection for Life-log System

Tremendous advances in wearable computing and storage technologies enable us to record not just snapshots of an event but the whole human experience for a long period of time. Such a "life-log" system captures important events as they happen, rather than an after-thought. Such a system has applications in many areas such as law enforcement, personal archives, police questioning, and medicine. Much of the existing efforts focus on the pattern recognition and information retrieval aspects of the system. On the other hand, the privacy issues raised by such an intrusive system have not received much attention from the research community. The objectives of this research project are two-fold: first, to construct a wearable life-log video system, and second, to provide a solution for protecting the identity of the subjects in the video while keeping the video useful. In this thesis work, we designed a portable wearable life-log system that implements audio distortion and face blocking in a real time to protect the privacy of the subjects who are being recorded in life-log video. For audio, our system automatically isolates the subject's speech and distorts it using a pitch-shifting algorithm to conceal the identity. For video, our system uses a real-time face detection, tracking and blocking algorithm to obfuscate the faces of the subjects. Extensive experiments have been conducted on interview videos to demonstrate the ability of our system in protecting the identity of the subject while maintaining the usability of the life-log video.

KEYWORDS: Video Analysis, Audio Analysis, Privacy Protection, Wearable Computing, Life-log System

(Jayashri S. Chaudhari)

(13th December, 2007)

Privacy Protection for Life-log System

By

Jayashri S. Chaudhari

Dr. Sen-ching, Samson, Cheung

(Director of Thesis)

Dr. YuMing Zhang

(Director of Graduate Studies)

13th December, 2007

(Date)

RULES FOR THE USE OF THESES

Unpublished theses submitted for the Master's degree and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgements.

Extensive copying or publication of the dissertation in whole or in part also requires the consent of the Dean of the Graduate School of the University of Kentucky.

A library that borrows this thesis for use by its patrons is expected to secure the signature of each user.

Name	Date

THESIS

Jayashri S. Chaudhari

The Graduate School

University of Kentucky

2007

Privacy Protection for Life-log System

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in the
College of Engineering
at the University of Kentucky
By

Jayashri S. Chaudhari

Lexington, Kentucky

Director: Dr. Sen-ching, Samson, Cheung , Department of Electrical and Computer

Engineering

Lexington, Kentucky

2007

Sagar

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere gratitude towards my advisor Prof. Sen-ching Cheung for his valuable guidance, encouragement and continuous support throughout this thesis work. It has been a wonderful experience in working in his research group not only in the field of research but also in terms of personal growth. Besides the technical guidance, I am very thankful for the financial support as well that I received throughout this work. Next, I like to thank other members of my thesis advisory committee, Prof. Donohue and Prof. Zhang for taking time to read my thesis and providing valuable comments.

I would like to thank my friends, especially the group mates for being very helpful and supportive to me. I would like to thank people at the Center of Visualization and Virtual Environment for providing excellent educational, conducive and healthy environment for students and also for participating in some of the experiments that I conducted to complete this work.

Finally, I am particularly grateful to my husband for his unremitting support and patience throughout this work. Without his love and encouragement this work would probably not have been completed.

Table of Contents

List of Tables

List of Figures

List of Files

1. JayashriChaudhariMSThesis.pdf

Chapter 1

Introduction

In this chapter we present objectives of this thesis, discuss the motivation behind our research, and also describe the major contributions of our work. First, we introduce the "Life-log" system which can record every experience of a person's life and discuss some of its potential applications in different areas. We also identify the technical challenges associated with the practical implementation of such system. Then we discuss the main focus of this thesis, the importance of *privacy protection in the life-log system* and how we can design and implement it for a wearable and practical unit. At the end of the chapter, we present an outline of the thesis.

1.1 Background

1.1.1 Life-log Systems

Memories form an integral part of a person's identity, therefore, every human being has an innate desire to capture experiences and preserve them. To this end, humans have successfully developed many technologies to record their memories permanently - handwritten diaries, letters, film cameras, audio recordings, digital images, right up to the most recent digital video. With the studies being made in wearable equipments and storage technologies, a significant and dynamic market of digital media has grown. The availability of portable multimedia recording devices allows us to record specific moment of our life, which naturally leads to a question "is it possible to design a device that can capture our experiences on the spot, and with details that a human

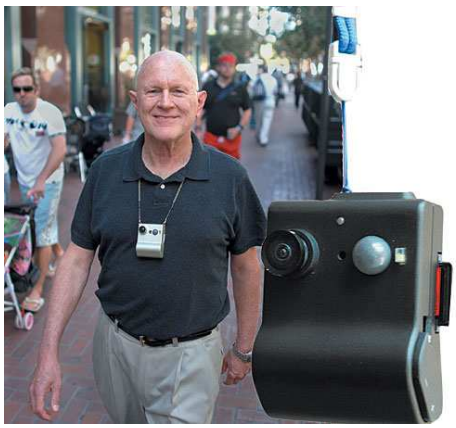
subject would feel them?". The concept of *Life-log system* is just that-, *It is a system that captures everything, at every moment and everywhere you go.* The vision of being able to record life time experiences has first been proposed by Vannevar Bush [1] more than fifty years ago. The interest in realizing this vision is still strong, as indicated by the amount of research conducted on the topic currently [2]. This wearable system will help the user in recalling experiences with search indices and provide significant details of the event just as the user experienced it. The technology is available now to allow us to build such a wearable system which can continuously record almost all human experience for days at a time. While there have been significant advances in the availability of hardware to build such a practical life-log system, the software needed to manage and analyze the huge amount of unstructured multimedia data that is captured is still lacking. In the following section we discuss some of the applications of life-log system and technical challenges associated with its implementation.

1.1.2 Applications of Life-log Systems

The Life-log system has applications in broad range of fields as shown in Figure 1.1. In this section we present a few of those applications.

Personal Archival System

The primary application of the life-log system is to record all memories of a person and organize them with different kinds of context information such as location, time, person's mood, occasion or any other relevant information that can help in retrieving an incident of interest. Life-log system can act as a



(a)

(b)

(c)

(d)

Figure 1.1: Applications of Life-log system (a) Personal Archival (b) Military (c) Law Enforcement (d) Hospital (Images are downloaded from images.google.com) (a)http://www.spectrum.ieee.org/images/nov05/images/recf1.jpg (b)http://www.openthefuture.com/images/853_web.jpg (c)http://www.apogeonline.com/webzine/2006/04/13/04/20060413040101.jpg (d)http://www.nist.gov/public_affairs/baldrige2004/RWJU_hospital_hires.jpg

memory-aid device in which the user can navigate through by using an intuitive interface and retrieve any event of his/her life with as much detail as possible.

Medicine

Another application is in medicine, where the stored personal records of a patient can provide multiple benefits. For example, the life-log system can provide

accurate details of each event that happened until a person fell ill (independent from the patient's memory which may not always be reliable). This will not only help the doctor in diagnosing the cause of the illness but will also increase the speed of the diagnosis. Learned life-patterns or habits can also provide clues to a doctor to predict potential medical problems. On the other hand, doctor also can wear the life-log system while examining a patient and the life-log system will take care of all cataloguing work related to medical reports. Again the system might provide past visits and all medical history of a patient very easily which will speed up the process of diagnosis.

Police Questioning

Life-log system has many advantageous applications in areas such as, on the (crime) spot police questioning and in events in which a crime witness/suspect is interviewed by the police in a criminal trial. Some people may feel uncomfortable in being interviewed at the police headquarter because of the tense milieu in the interrogation room, and would prefer to give their testimony in their own home where they are more relaxed. The life-log system can be useful in recording the testimony in such situations because it is inconspicuous and extremely mobile [3].

Law Enforcement

Since the 1990's submitting the evidence of a crime in court proceedings in video format has been increasing and found to be extremely useful. In the past, videos were captured in VHS format with the video capturing system installed in the

police patrol car. With many advances in digital technologies, nowadays the evidence is recorded in digital format, which by itself offers many advantages over the VHS format [4]. However, the system still has disadvantages due to the fact that the the video capturing system is fixed in the car. This severely restricts the area that can be captured and also requires that the system be manually controlled. For example, during a police chase, one additional police officer (besides the driver) will be needed to continuously focus the camera on the suspect, and to adjust the zoom to capture the video with sufficient detail. In such situations the life-log system can be very useful. When a police officer mounts the camera on their body while he/she is on duty, the officer can capture evidence at a crime scene, on their own, and without much interference or restriction on their movements. Because the officer will be able to take the camera everywhere, adjustments to the zoom are not needed (or are minor at best) in recording quality video. This advantage can get rid of the manual adjustments needed for a fixed camera and also the need for a dedicated person, which can greatly increase the efficiency of the officers. The other advantage is that as the life-log system is completely mobile, the capturing area is now unrestricted and it is possible to record evidences in places where the current on-patrol camera system is unable to reach [3].

Military

In the military, soldiers are often required to report all observations or incidents with sufficient accuracy and depth after returning from patrol/combat. The life-log system can be very useful in such cases, because it can improve the reporting

capability of soldiers by providing a record of the event with enough detail and accuracy. The system would record information such as the location, audio, video, motion and other data via body worn sensors, which would be then used to help write the report. The vision for such a system, for example, has been proposed in a DARPA research project called "Advanced Soldier Sensor Information Systems Technology" (ASSIST) [5].

1.1.3 Technical Challenges

The design of a life-log system to store and manage a person's life time experiences poses many technical challenges in computing research. We discuss some of these challenges in this section.

Information Management

How can we manage and store automatically the inordinate amount of data that is being recorded in an effective way? What kind of supporting information such as annotations or other contextual information be stored along with raw data and what means should be used to collect those data? The challenge is in identifying ways to integrate all the different modes of information captured through various sensors and effectively organize the collected data, and to interconnect the related data.

Information Retrieval

Solely capturing and storing the memories is not sufficient. In order to be useful, a system like the life-log should provide an environment which will facilitate browsing of stored experiences. The question of interest is, "How to

index/summarize the recorded data to be able to retrieve useful or relevant information easily and effectively?". How to design an intelligent query processor which can perform sophisticated interpretation of stored data? For example, one may ask the query: "Search all vacation videos at Grand Canyon while hiking down to the Colorado river". Existing solutions such as manually cataloguing or annotating every segment of life-log video would not easily scale to the large amount of data collected by a person over a life time.

Knowledge Discovery

Knowledge discovery concerns with developing methods to automatically analyze the collected information and discover important information. Our day-to-day lives involve many repetitive events such as having breakfast in the morning, driving to office, checking emails, coming back to home, play with kids, watch TV, go to sleep, etc. From the analysis of the life-log video, a life-log system should be able to learn the life-pattern or specific traits or habits of the user of the system. Such analysis can be very useful in the medical diagnosis of a person because it provides the doctor with important information regarding the person's lifestyle. Designing such an intelligent system would require the combination of many aspects of artificial intelligence, such as speech processing, context recognition by audio-video analysis, object recognition, natural language processing, machine learning, etc.

Human Computer Interaction

Developing an effective user interface to be able to interact with the collected

information is crucial for the utility of such a system. It is also imperative to learn how to present the collected information and associated knowledge intuitively to make it useful to the user. Solution of these issues requires that the system be capable of learning from the collected data.

Security and Privacy

Another aspect for the implementation of the life-log system are in ensuring that the privacy of individuals captured in the video can be protected and that the collected data is safely stored with necessary security provided. There are many technical and legal policy issues needed to be addressed before the life-log system can actually be deployed. For example, if a person(A) is present in other person(B)'s video, what rights does person(A) holds on the video with respect to its distribution or use? Can the system selectively block out those who do not want to be filmed? Will such privacy protection scheme diminish the usefulness of the video? Therefore, the challenge is in developing technological solutions to protect the privacy of people according to their needs. The system should also be able to provide a solid security mechanism so that if the collected data falls into wrong hands, it should not be misused. The goal of this thesis is to provide practical solutions to privacy issues of subjects who are being captured in the life-log video. In the next section we further elaborate on the importance of ensuring privacy in the life-log video.

1.2 Motivation

1.2.1 Importance of Privacy Protection

Privacy is one of the most important issues that need to be carefully handled in a system that records everything, everywhere, and at every moment. Although every new technology like the life-log can provide benefits to society with proper usage, it can also prove harmful if misused. Therefore, people may feel reservations/fear towards system like the life-log due to the possibility that the audio-video recording of naive individuals could be misused. The recording of life-log videos also threatens a fundamental right to privacy of every individual in the United States. The bill of rights of the US constitution protects the right to privacy of citizens, although the word "privacy" is not explicitly mentioned anywhere in it [6]. Recent advances in technology seem to threaten this very basic right. For example, the prevalent video surveillance and sensor networks in the country makes everyone feel like they are being captured and observed every moment. This makes people more wary and their attitude towards new technologies becomes more negative, which ultimately hampers the progress in computing research [7]. Some of the emerging technologies with privacy concerns are face recognition, biometrics, video surveillance, sensor network, semantic web, bio-terrorism surveillance, etc. Despite the potential benefits that these technologies offer to the society, many citizens are concerned that their privacy will be invaded and this forces people to choose between their own privacy or benefits of the technology. For example, consider face recognition technology which was tested by police during Super Bowl game at Tampa, Florida in January, 2001. To supervise all

the people coming to watch the game in the stadium, police installed face recognition system to capture faces of people and compared them to the database of criminals. The event was strongly criticized by many privacy advocating organizations such as American Civil Liberties Union; who argued that the public was not made aware of the system which invaded their privacy rights, and that the system could make false recognition which makes it less reliable [8]. Due to the resistance to this privacy threatening technology, in August 2001, Florida City Council passed a legislation to ban the use of face recognition technology by police officers [9]. In another incident, a famous DARPA project called Life-log was attacked in a similar way and was later aborted [10].

Another obstacle in the implementation of these technologies arises from the problem that the term *"privacy"* is vague and subjective. Different people seem to have different level of privacy concerns or even different concepts of privacy. Gathering "private" information may consists of activities such as recording or listening what a person is speaking, knowing his/her location, identifying his/her activity at any given time. Intrusion of privacy also include the manipulation of any data (eg. audio or video recording) belonging to another person. This gives rise to the question of what exactly we should preserve to protect privacy. Also the rules and regulations about multimedia capturing are not uniform across states in the United States. Every state has different rules about what is legal to record and what is not, and also what rights a person has over another person's recordings if he/she is present in it. These are some of the issues which illustrate that privacy is not well understood. The book "Digital Person" [6] provides an excellent discourse of what constitutes "privacy"

and how laws can be designed to balance the need to protect privacy in the current information sharing age with the benefits that new technologies bring to individuals and the society at large.

From the above discussion it is clear that it is extremely important to first provide technological solutions which ensure that an individual's privacy is protected. Only after that the deployment of such technologies is possible. In this work we address the privacy issue in the life-log system by providing reasonable solutions to hide/mask the identity of subjects recorded in the life-log video. The next section explains the major contributions of this work.

1.2.2 Contribution of the Thesis Work

Legal and privacy related issues in a life-log system called *Total Recall*, which captures all experiences of a person has been previously discussed [11] but there was no implementation. In addition, prototype systems for privacy protection in video surveillance [12, 13] and in face recognition [14] are also available. However, to the best of our knowledge, there has been no automatic solution that protects privacy information in life-log video recordings.

In this work we present a practical real-time privacy protection mechanism for subjects captured in a life-log video. Our work mainly focuses on interview scenarios where producer (who is the user) of the life-log videos is interviewing one subject in a relatively quite room. Such scenarios can occur in many applications such as police questioning in police headquarters, a journalist interviewing a person, and doctors examining a patient etc. We use *real-time face blocking* and *voice distortion* to conceal

the identity of the speaker. We have decided to block only face instead of removing the person all together to keep the video useful to the producer, because body language of the subject is conveyed through video with minimal disclosure of subject's identity. Voice identity of the subject is protected by voice distortion achieved by pitch shifting of audio signal. There are two major contributions of this research work. First, we have implemented real time subject's face blocking and subject's voice distortion to protects his/her identity. Second, we perform detailed analysis on the audio distortion algorithm for its performance in concealing the speakers identity, and also the degree of ambiguity that ensures that no distinctive features remained in the distorted signal, similar to k-anonymity [15]. To keep the recording useful after audio distortion, we measured the subjective intelligibility of the distorted conversation. Neither the face blocking algorithm and the pitch shifting algorithm are novel invention. While there have been evaluation studies on visual privacy protection [16], we are not aware of any studies on audio privacy protection. In Chapter 5, we provide a detailed study on how well a pitch shifting algorithm can protect individual audio privacy while maintaining the usefulness of the audio signal. This is a novel study, which, to the best of our knowledge, has not been conducted before by any other group.

Our experiments show that a good balance between privacy protection and usability can be obtained by implementing our scheme. The pitch-shifting algorithm achieves 100% speaker identification error and thus allow perfect audio privacy protection. To determine if the recording is useful after distortion, we compare intelligibility of the conversation before and after the distortions in terms of Word Error Rate(WER). The average WER before and after distortion (with $\alpha = 1.40$) are 14.2

and 14.4 respectively. Statistical analysis shows that the clarity of the recording for pitch scaling factor $\alpha = 1.40$ is same as that of the original audio clips. Analysis of segmentation algorithm gives recall between 0.86 to 1 in detecting the subject speaking portion of the audio.

1.3 Thesis Outline

The thesis is organized as follows. After providing the background and stating the major contributions in this chapter, Chapter 2 gives a comprehensive summary of the existing systems and research work related to personal experience capturing and archival technologies. It provides a survey of the research efforts in fields such as wearable computing, multimedia processing, and information retrieval. It also summarizes the research work related to privacy protected information sharing and its application toward the realization of life-log system. In Chapter 3, we discuss the system architecture of our life-log system. We explain the hardware and software architecture of the life-log system and give a detailed system overview. In Chapter 4, we explain the privacy protection mechanism implemented for audio-visual information. First the design objectives are discussed and then an explanation of the different algorithmic components such as face detection and blocking, audio distortion by pitch shifting and audio segmentation method are given. In Chapter 5, we present the experimental results from the tests conducted on our privacy protection scheme and also evaluate the scheme's performance. In the final Chapter 6, we summarize the conclusions from the research work and provide suggestions for future work in this area.

Chapter 2

Related Works

This chapter is divided into two sections. In Section 2.1 we present the existing research work related to the life-log technologies, while Section 2.2 discusses the current research efforts towards privacy protection mechanisms in various fields including audio-video capturing.

2.1 Experience Capturing Systems

In the article "Capturing Experiences Anytime, Anywhere", the authors present an excellent overview of the chronological advancements in mobile-experience capture technology [17]. The paper discusses how the present Web logging culture may be supplanted tomorrow by a "Sensecam", a wearable camera, which utilizes sensors to detect interesting events and triggers the capturing of images [18]. In the past few years, research in wearable computing, video retrieval and databases to record a person's life time experience has been steadily growing. A pioneer in wearable computing, Steve Mann, has experimented with wearable cameras as a means for recording personal events. But he did not give sufficient consideration to how the abundant amount of collected data can be organized and utilized to extract information [19]. An ongoing Microsoft research project "MyLifeBit" also attempts to realize the Memex vision of V. Bush by digitizing a person's life via exhaustively recording his/her experiences through documents, emails, books, web pages the person visited, and digital photos and videos [2]. Research Scientist, Gordon Bell, has been recording

his whole life into digital format and has been experimenting with the collected data. Similarly, Lamming and Flynn, used the ParcTab system [20] to design a portable episodic memory aid called the Forget-Me-Not system [21]. They used various sensors that were implanted in a laboratory or badges worn by users to record contextual information such as location, encounters with other individuals, personal activities, file exchanges, and workstation activities to capture and retrieve different life experiences. Clarkson developed a system with fisheye video cameras that can be mounted on a chest strap from the front side. He used the system to continuously record his day to day experience for 100 days [22], and experimented with the collected data. Jennifer et al. has developed the "Startle Cam", which records video data triggered by the user's physiological reactions such as skin conductivity, heart rate, respiration rate and muscle activity [23]. Finally, in the wearable computing area, sensors such as acceleration sensors, GPS, physiological sensors (brain wave analyzer), skin conductivity and gyro sensors have also been used to trigger the capturing of an event and retain its context [24].

Ubiquitous experience recording technologies is an emerging field that has grabbed significant attention from the research community. From 2004 ACM has been conducting a workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE), devoted to technological topics of capturing life-time personal experience, sensors, and related research challenges [25]. Pervasive 2004, the second International Conference on Pervasive Computing included a workshop on Memory and Sharing of Experiences. The UKCRC (United Kingdom Computing Research Committee) has identified *Memories for Life* [26] as a significant challenge in computing research [27].

DARPA's (Defence Advanced Research Projects Agency) ASSIST program [5] actively funds research for developing body sensor technology which the soldiers can wear to improve their capability to recall and report observations and experiences during ground duty.

The existing research work related to life-log technologies is focussed on information summarization, indexing and retrieval issues, and to manage life-long collection of personal information. By taking into account the fact that the human mind tends to remember an experience by associating it with the context of the situation, most of the research in Life-log video archival systems has centered around making use of contextual information such as location, time etc. captured by the additional sensors attached to the system. Previously Aizawa et al. and Hori et al. have shown how a person's brain wave and motion captured from sensors (which record human emotions) can be used to summarize the life-log video [28, 29]. In another paper, [30] the authors have used both content and contextual information to extract keys/meta data to annotate the video sequence. They discuss about detecting a conversation scene by analyzing audio signal and face detection technology. Time-constrained K-mean clustering technique has been demonstrated [31], in its ability to learn the underlying structure of continuously recorded personal memory archives. Some studies have focused on using continuous audio-only recordings because it may offer advantages [32]. The authors used spectral clustering techniques [33] to segment and cluster the audio recording into different 'episodes' that can be related to the context of the situation experienced by the user. The author calls the user context as *acoustic environments* which includes different locations, different user activities, etc. Each of the envi-

ronment has their own acoustic characteristics which is useful for segmentation and schematic labeling. The authors argue that making an audio recording requires minimal attention from the user because the sensitivity of the microphone is not affected significantly by its positioning and angle, and it can be easily carried with little inconvenience. In another paper, the same authors study various audio features that can be used to represent one-minute audio window, and their effectiveness in segmenting audio recording [34]. This has also triggered research work related to context recognition through content i.e. audio, video information. In [35], continuous audio-video recording obtained by a wearable camera and microphone mounted on chest strap was classified into different events such as, entering office, leaving office etc. Hidden Markov Model (HMM) was first trained by manually labeling the stream with event names by the user as they occur. The trained HMM was successfully able to classify the recording into twelve different events. In [36], the authors describe object and place recognition method which make use of context information extracted from video frames captured by a head-mounted wearable camera. The paper describes an algorithm using HMM to classify video segments by their background environment into categories such as office, kitchen, etc.

These are the major research work in continuous personal archival systems. In the next section we discuss research work related privacy protection, faced by these experience capturing technologies.

2.2 Privacy Protection

Despite the recent research efforts in indexing and summarizing life-log content, we are still a long way from fully achieving a practical life-log archival system. Privacy and legal issues related to it are largely neglected and need to be tackled carefully. Here, we discuss some of works to address privacy challenge in different domain.

Public data Release

There has been significant research work regarding privacy protected data mining in context of public data release [37] [38] [39] [40] [41]. Many public, and private organizations, such as hospitals, banks, and government agencies constantly collect personal information of people. They are required to release/share these microdata which is generally in the form of relational tables, for different statistical studies such as demographic analysis, public health research, etc. The challenge that these organizations face is how such data can be released *anonymously* i.e. in such a way that the subjects of the data can not be identified, thus ensuring their privacy; but at the same time the released data should remain useful for practical purposes [42]. Anonymity and privacy in microdata release has received tremendous attention from the research community, and there has been significant efforts towards making the released information anonymous [43] [44] [41] [38] [39]. One of the approaches discussed in the literature is to use perturbation techniques to obfuscate the actual data [45] [40]. Another approach is to suppress or generalize some of the sensitive information [41], [46].

Although released microdata is made anonymous by removing identity of a person, it raises serious concerns about privacy of people, because a person can be re-identified

by relating the information gathered from different sources [46]. Privacy protection model called K-anonymity, has been proposed to reduce the risks of such correlation attacks to re-identify the person [15] by combining information from different sources. The authors explain the definition of k-anonymity as,

> *"A table satisfy k-anonymity if every record in the table is indistinguishable from at least k-1 other records with respect to every set of quasi-identifier attributes; such a table is called a k-anonymous table".* The quasi-identifiers are the collection of attributes in a table that can be linked with external data to uniquely identify individuals in the population.

K-anonymity model has shown good results in recent years and has been widely used. In [47], the authors extend the definition of k-anonymity and propose a model which can be applied to data mining algorithms i.e. data mining algorithm using the proposed model will always generate k-anonymous data.

Image or Video Data (Video Surveillance)

While there are well established privacy models for microdata, there is yet no formal privacy framework for multimedia data in the literature. Multi modal nature of multimedia information makes it harder to achieve privacy protection in it. Due to the richness of information carried by multimedia data, it becomes more difficult to gain anonymity.

Algorithms have been presented to de-identify images in surveillance video to address the privacy protection issue [48] [49] [50] [16]. Most of the techniques use simple obfuscation methods to filter out any sensitive portion of an image. One

common method is to pixelize or to blur sensitive areas such as a person's face in the image [48] [51]. Hudson and Smith [52] presents a shadow-view filter that compares the static background image of a scene and the video frames block-by-block basis of 8x8 pixels. If the difference exceeds a threshold, the block is replaced by the average intensity pixel of the block in the background image. This gives rise to a ghostly shadow like visual effect of foreground objects in the video. Dufaux et al. [50] presents a method using transform domain scrambling of privacy-sensitive portions in the video surveillance to address privacy protection. Discrete Wavelet Transform (DWT) coefficients corresponding to the sensitive portion of the video, are scrambled by randomly inverting their signs. The advantage of this method over the other obfuscation methods is that it is reversible. Private encryption key is needed to unscramble the video and thus it provides sufficient security. Some methods block out the sensitive information in a video stream [53] [54]. privacy filters, defined by privacy grammar, are used by privacy buffer that operate on incoming video signal to filter out sensitive information [53].

A. Senior et al. [55] discuss the definition of *video privacy* in detail and how it differs in comparison to privacy in other data. Authors also describe privacy protection model which is based on object oriented representation of a video scene. Depending on access control level of an individual, the system re-renders the modified video with privacy-sensitive portion blocked or replaced by computer graphics. The implementation of these concepts is called "PrivacyCam" by the authors, in essence it is a smart camera that produces privacy protected video signal with sensitive information removed. In another research article [56], the authors describe a distributed privacy

paradigm for Visual Sensor Networks (VSN). The author defines VSN as a sensor network in which visual data is captured along with sensor data, that can be useful for improving the service provided by sensor networks when deployed in a hospital, combat field, etc. The privacy and security concerns are high in VSN, which requires captured visual data to be properly secured to prevent illegitimate attacks. The paper describes TANGRAM algorithm based on Lyapunov stability theory, which enables sharing of images among sensor nodes which modifies images in such a way that if illegitimate attackers get access of images they cannot break the puzzle to get original images, thus achieving security and privacy.

A more formal model of privacy called *k-same* that specifically targets face images is presented in [14]. A face de-identification algorithm is proposed in which many facial characteristics are retained but face recognition software can not reliably identify the face. The algorithm is based on defining similarity map between faces and creating new faces by averaging out image features of most similar faces. This algorithm implements k-anonymity protection model for face images. The excerption of definition of k-same taken from the paper which succinctly explains it is as follows,

> *"Given a person-specific face set H; and, a face set H_d which is k-anonymized over H using a preserving face de-identification function* f:$H \rightarrow H_d$, *if* f *is effective with respect to the claim:*
>
> *Given any face image $\Gamma_d \epsilon H_d$, where $\Gamma_d = \mathrm{f}(\Gamma)$ for $\Gamma \epsilon \boldsymbol{H}$, there cannot exist any face recognition software for which the subject of Γ_d can be correctly recognized as Γ with better than 1/*k *probability.*
>
> *then* f *is a* **k-Same de-identification function** *and H_d is a* **k-Same**

de-identification. *The goal is to determine the appropriate function* f *with minimal information loss."*

Lack of Evaluation Techniques

Despite having a number of techniques for privacy protection, there is no formal evaluation model which can be used to measure the degree of assurance of privacy provided by each technique. One of the few research works in this area is a paper by Boyle et al [48], in which the authors study the impact of filtering methods, pixelation and blurring, on awareness and privacy in *media spaces* i.e. video signal. Media space application requires sharing of video across offices and rooms to provide *informal awareness* to distant work-groups working in an organization [57]. The experiments involved two subjective studies with modified videos of varying degree of filters applied to it. First, the test subjects were asked to extract awareness information such as number of people in the scene, their activities, their gender, and other visual information. In the second study, each filter level was tested for its ability in preserving privacy by asking test subjects to identify a person in the filtered video. In another paper [16], Zhao and Stasko evaluate various image filtering based techniques in media space application for their performance in providing enough awareness information while preserving privacy. The authors first describe image filters such as blurring, pixelization, edge-detection, shadow-view, live-shadow filters to hide sensitive regions in a video and ensure the privacy of people. They also report experimental results that elucidate in a qualitative and quantitative manner the degree of awareness achieved by these filtering methods.

Audio data

Besides visual identity, a person also has an audio identity. Our speech signal carries privacy sensitive information, such as a unique speech pattern that identifies an individual and the meaning of the speech i.e. what is being said. There has been lots of research work in the voice conversion domain [58] [59] [60] [61] to modify the audio identity of a person by changing the voice pattern of one person to another while preserving the speech content i.e. meaning of the speech. The methods discussed in literature involve changing physical properties such as voice track pattern or pitch that encodes an identify of a person. In [62] the author describes different methods, such as linear prediction, cepstral analysis, and pitch alteration, that are used for voice transformation. In [58], the authors present a voice conversion method based on vector quantization and spectrum mapping. The method produces a *mapping codebook* which shows correspondences between codebook of source and target speaker. Another paper [59] describes a voice conversion method based on probabilistic conversion between source and target spectral envelope modeled in Gaussian Mixture Model. Most of the research work in voice transformation has focus on the applications such as entertainment, in movies etc. But application of voice conversion for privacy preserving technologies and security has not yet been explored very well. Voice transformation can be used to address the threat to privacy brought by speaker recognition techniques [63] [64].

Chapter 3

Proposed Life-log System Design

In this chapter we describe the hardware and software architecture of the proposed life-log system.

3.1 Hardware Design

Among the most important design criteria of wearable life log system from the user's point of view are that the system should be comfortable, light-weight, and should not interfere with daily activities. The system should be able to capture both audio and video data for long duration and provide an intuitive user interface to review and edit the data. To apply this system in the law enforcement domain, wireless connection capability is also important, as it allows the police to receive critical updates about criminal activities from the central command.

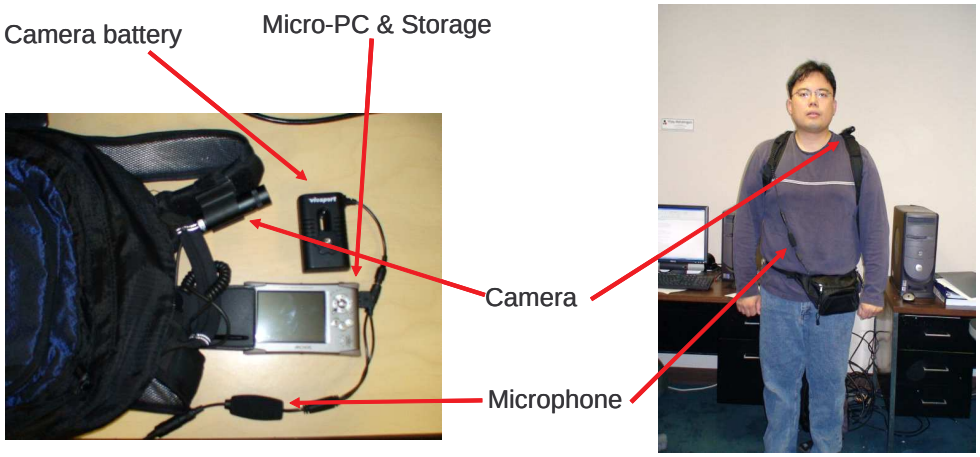


Figure 3.1: Main components of the proposed wearable life-log system: a small camera mounted on the shoulder, a microphone, and the processing, storage and browsing unit in the small backpack.

Based on the above criteria, we propose the design of a wearable life-log system as shown in Figure 3.1. It consists of three main components:

Small Helmet Camera

The suitable camera should be of small-size, light-weight, and weather-proof. It should have high spatial resolution, high frame rate, high low-light sensitivity and flexible mounting options.

In order to obtain a clear, stable and unobstructed view of the scene, an appropriate mounting position for the camera is crucial. Most of the existing wearable systems have their cameras attached to a helmet or the spectacles, or are directly mounted on a head-band [65]. Although this position captures the scene most accurately as seen by the eyes, the camera requires a helmet or special spectacles both of which are undesirable because they easily draw the attention of the surrounding people. Also, due to the inadvertent movement of the head, the video captured from this position is highly unstable. Therefore, we propose to secure the camera on the shoulder. Our preliminary experiments show that the video is far more stable than that obtained from a head-mounted camera. A lens with a wide field of view (at least 70^o) should be used in order to capture most of the frontal scene. Based on our research, a good choice that matches our criteria is the S.C.O.U.T. camera from Viotac shown in Figure 3.1. This camera weighs 105 grams and captures NTSC resolution video with a field of view of 72.5^o and light sensitivity of 0.2 Lux.

Omnidirectional Microphone

The primary function of the microphone is to capture the voice of the user of the wearable system as well as those who have face-to-face conversations with the user. This requires that the microphone not be too close to the mouth otherwise the user's own voice will be too loud. The omnidirectional microphone that comes with the S.C.O.U.T. system is suitable for our purpose. We can secure it to the middle of the front strap of a small backpack that holds the rest of the wearable system. The microphone is then roughly at the heart position of the user which is just far enough from the mouth.

Processing Unit (Viola Micro PC)

If the life-log system is only used to store the video without any form of analysis, it is sufficient to use one of the many personal Digital Video Recorder (DVR) devices in the market to store the captured content. The most popular model, the Apple iPod, can store up to 150 hours of video and it weighs less than 6 ounces. Nonetheless, this kind of DVR (Digital Video Recorder) has very limited user interface support and it is almost impossible to program such a device. Even though we do not expect the user to constantly interact with the system, an adequate user interface should at least support multiple windows and point-and-click functionality. For example, the user may want to find all the scenes that contain a particular object detected by the system.

On the other hand, a laptop is too heavy and bulky to be practical, and a Personal Digital Assistant (PDA) usually does not have enough memory for media storage. As a compromise and to keep the weight of the overall system

minimum, we have chosen to use Sony's VAIO Micro PC, which weighs just 1.2 pounds. It is small enough to fit in the palms of your hands with the size of 6 x 3.74 x 1.27 - 1.5 inches, it has a WSVGA touch screen XBRite display, and it runs Windows XP operating system with 1.2GHz processor speed and 30GB memory. It also has WiFi 802.11a/b/g, Bluetooth, 10/100 Ethernet which is useful for wireless connectivity and can provide more localized position tracking. Again it has biometric fingerprint scanner which is very useful to provide security if in case the user looses the device. It does not have in-built GPS unit, but can be added easily which can provide location information to be used by the life-log analysis. This is important, as it allows us to process the video in real time to run the privacy protection scheme in Micro PC. The fully charged Micro PC can capture up to 5 hours of data continuously.

The processing unit will perform operations such as extracting key frames for indexing and summarization, privacy protection, and provide a user-interface to interact with the user. This processing unit serves three functions viz. it provides an intuitive user interface for browsing, provides sufficient storage space to store one day's worth of video and finally it provides processing capability for initial manipulation of incoming video.

The weight of the entire system is about 2.0lbs. The Micro PC, and other accessories are most easily carried in a small back pack strapped on the shoulder as shown in the figure 3.1. Due to the constraints on weight and power consumption, the wearable system can only perform some initial analysis tasks that provide a quick

feedback to the user. More sophisticated processing needs to be done off-line after the video has been transferred from the Micro PC to a desktop computer.

3.2 Software Design

In this section, we describe the software design of our proposed system which includes the software architecture and the algorithmic design of its components.

3.2.1 Software Architecture and Basic Components

Figure 3.2 shows the software architecture of our proposed life-log system. It consists of five main processing components and two databases. The processing components are Raw Data Capturers, Feature Extractors, Change Detectors and Object Detectors, Privacy Protectors, and View Generators. A processing component is a high-level behavioral description of a particular kind of input/output processing. In the sequel we will describe some of the specific algorithms grouped under each unit. The two databases are Raw database, and Meta-data database. These databases provide support for simple indexing and browser interface.

Raw Data Capturers

The *Raw Data Capturers* unit is the lowest layer of the software architecture and is responsible for interacting with the hardware and obtaining the raw audio, video, and location information.

Feature Extractors

The *Feature Extractors* unit removes noise from the audio and video and extract various types of attributes that are amenable to analysis. Examples of audio

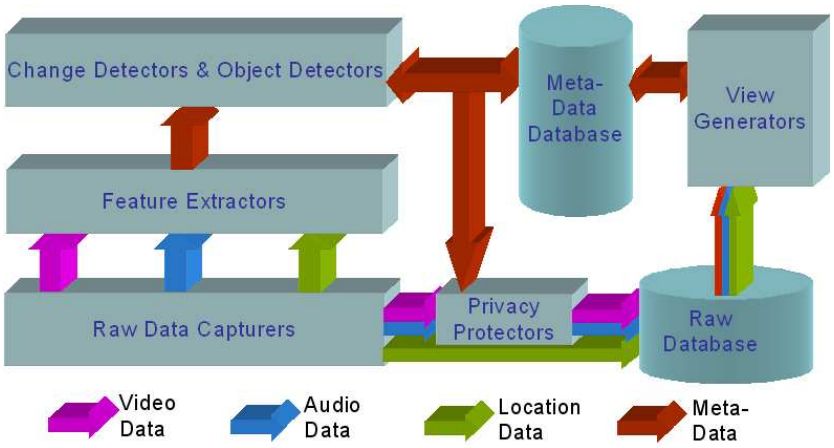


Figure 3.2: Software architecture of our proposed system.

features include short-term loudness, frequency spectrum, pitch, and timbre. For video, there are three types of features: global, local, and temporal. Global features such as color and edge histograms are useful for characterizing the physical environment. Local features such as various types of corner detectors are useful for object identification. Temporal features like motion vectors are useful for characterizing temporal events like walking or running. These features are fed to various *Change Detectors and Object Detectors*.

Change Detectors and Object Detectors

A change detectors unit builds an online statistical model of various features and reports the time instance when there is a significant change in the model. It is useful for partitioning the video in temporal dimension into logical segments which can be much more efficiently manipulated than individual video frames.

Recently, there have been tremendous successes in combining many simple features in detecting specific objects such as faces and license plates [66]. Audio

Figure 3.3: Privacy protected video frame.

features can also be combined to identify various physical locations [67] and audio events, like conversations [24]. After appropriate training, these object detectors are computationally efficient and thus are appropriate to be implemented in the wearable system. The face detector will also be used for the *Privacy Protectors*. When the privacy protection mode is on, the portions of the audio that contain voices other than the user's voice will be distorted and all the faces detected in the video will be blocked as shown in Figure 3.3.

View Generators

The *View Generators* is used to generate various user-interface views in support of various types of browsing. The main goal of view generator is to provide a

quick and useful summary of the recently-captured video so that a user can easily browse through a day's worth of data

Privacy Protectors

The *Privacy Protectors* provides the real-time privacy protection mechanism which filters the incoming video for identity information of subjects being captured. This thesis work mainly focuses on this module. We give a brief overview of this module of the life-log system in section 3.2.2.

Raw Database

The raw database consists of raw data collected by the life-log system. The raw data comprises of the unprocessed audio-video data recorded, location information provided by the GPS system, time-stamp information and other data collected from different sensors attached to the system. The collected raw database is processed by *Change Detectors and Object Detectors* off-line and they extract different annotations for life-log video summarization purposes.

Meta Database

The Meta database consists of annotation data such as key-frames to represent fragments of the video, contextual labels such as location, time, etc. extracted from *Change Detector and Object Detector*. This database is used by the *View Generators* to provide meta data to create user interface.

3.2.2 Privacy Protection Module

We introduce two terms here, *Producer* to describe the person who wears life-log system, and *Subject* to describe any person besides *producer* who is being recorded by the life-log system.

We propose a real time subject's voice distortion and face blocking approach to address the privacy issues. When the privacy protection mode is "On", the face detector will detect and block the subject's face before it is stored in the life-log video as shown in figure 3.3. Besides the face, the subject's voice could also expose his/her identity easily. Therefore, only face blocking does not guarantee privacy protection. To disguise the subject's voice, the proposed system distinguishes the subject's voice from the producer's voice and distorts it so that the subject can not be identified.

Some of technical challenges while implementing the privacy protection scheme are accuracy, speed, and selectivity. Accuracy: It is paramount to achieve 100% accuracy in face detection because even a very small inaccuracy might disclose the person's identity. The problem is complicated by the fact that all the face detection algorithms discussed in the literature thus far only work for front face view. Audio distortion would require the analysis of the audio signal, and in the event of a noisy environment the subject voice detection accuracy can be reduced. Speed: It is highly desirable that the protection mechanism work in real time. This poses challenges in improving the speed of face detection and voice distortion algorithm and in reducing its complexity. Selectivity: Some people may mind being recorded by the life-log system. To handle such cases, the system needs to be designed such that it can distinguish those two groups of people. The selectivity problem will not be addressed in this thesis work. We

are dealing with one specific scenario of one-on-one conversation between the producer and a subject, and hence the issue of selectivity does not arise. For example, a case in which a single crime witness/suspect is interviewed by the police officer. This scenario is also applicable in situations such as in job interviews, or in hospitals when a doctor is examining a patient etc.

A simple method to detect the subject's voice by analyzing the power of the incoming audio signal is proposed. Because the microphone of the life-log system is closer to the producer as compared to other people around him/her, the signal power is significantly larger when the voice comes from the producer as compared to the subject. By using Adaptive thresholding, the method tries to classify the audio sequence into three categories viz. ambient noise or silence, producer speaking, and subject speaking. The subject speaking portion of the signal is then distorted before it is stored in the raw database. Voice distortion is achieved by altering the pitch by the Pitch-Scale Synchronous Overlap and Add (PitchScale SOLA) method discussed in [68]. The distortion is accomplished in two steps, first by time stretching and then again re-sampling it to make it of the original length. One problem with this method is that it might work poorly in a noisy environment. Further improvement in the method can be achieved by filtering out the noise and then applying the algorithm.

For face detection, we have decided to implement the algorithm discussed in [66]. The algorithm is based on image representation called "Integral Image" which allows fast calculation of Haar like features that are used to detect faces. The method works for frontal face views with good detection accuracy and can detect faces rapidly and is therefore suitable for real time applications. Another algorithm [69] that

could potentially be suitable for rapid face detection. The algorithm from [66] is further improved by combining face color tracker with face detector to deal with non-frontal face view. Basically, once the detector has recognized a frontal face, the face color tracker will ensure that the detector will continue to recognize the face even after it is no longer a frontal view. A similar method is presented in [70] which combines face detector and color based object tracker called PCI (Pixel Classification and Integration). The next chapter 4 *"Audio-Video Privacy Protection Scheme"*, explains the privacy protection mechanism that we have implemented for audio-visual information.

Chapter 4

Audio-Video Privacy Protection Scheme

In this chapter, we discuss our audio-video privacy protection scheme for life-log videos. Our current design focuses primarily on the interview scenario, in which the producer, i.e. the person who is wearing the life-log system, is speaking with a single subject in a relatively quiet environment. Interview scenarios are very important, and have its own specific characteristics which differentiate the interview video from the other life-log videos. In an interview video, there are not many foreground objects and the background stays almost always unchanged. Such interview scenarios appear in many situations, such as in hospitals where doctors are examining patients, in a police interrogation room where a witness/suspect is interviewed, or even a job interview recording. Thus, the simplicity and wide applications of the interview scenario are attractive for our research objectives and therefore we decided to implement it in this work.

4.1 System Overview

This section discusses in detail the privacy protection module introduced in Chapter 3. We first discuss the technical goals that are to be achieved and then explain the design of the privacy protection scheme to achieve those goals.

4.1.1 Design Objectives

There are four main objectives in implementing our privacy protection mechanism.

1. Privacy

   The main and obvious goal of the scheme is to protect the privacy of subjects who are being captured in a life-log video.

2. Accuracy verses Usefulness

   The protection scheme should provide enough information to make the data useful for review. For example, in Figure 4.1, the first image conveys all information but does not protect the person's identity, while the second image essentially removes all information but perfectly hides the identity, subsequently losing its usefulness. The third image tries to keep the balance between privacy and usefulness. Instead of removing all information, we can just block the face, because by blocking the face we can achieve significant privacy protection while maintaining the reasonable usefulness of recording. The other visual information such as body language of the person can be useful to relate to the conversation, making it look more natural.

Figure 4.1: Accuracy Vs Usefulness

3. Anonymity or Ambiguity

   Another important criterion is called k-anonymity where k is an integer measuring an ambiguity created in the data after privacy protection. The k-anonymity is formally defined in [15]. For example, a n-anonymous *(k=n)* privacy protection scheme distorts the data set of n distinct records of n different persons, in such a way that every individual will look and sound identical. Even if an attacker knows the identity of the subject in one video, he/she will not be able to gain any knowledge about other videos. In practice, it is difficult to achieve n-anonymity especially in multimedia data and thus the goal is to make k as large as possible and close to n.

4. Speed

   It is highly desirable that the protection mechanism be fast enough to work in real time. The small physical size of a life-log system limits its computational power and thus it is imperative to design highly optimized algorithms.

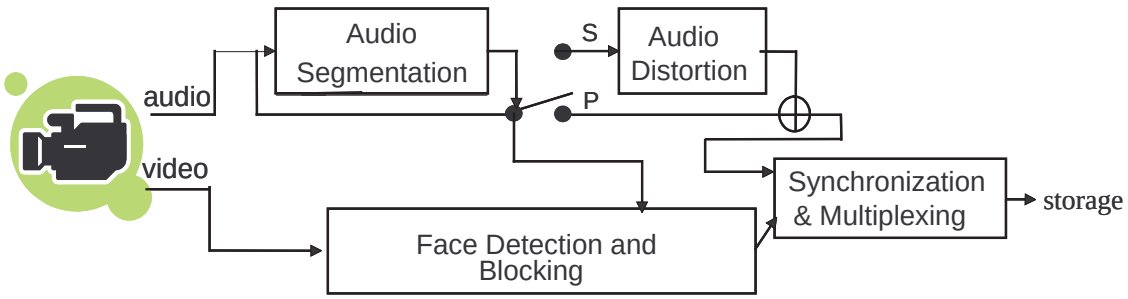
4.1.2 Privacy Protection Scheme

Considering objectives described in the previous sub-section, we present the overall design of the privacy protection scheme of our life-log system. The multi-modal nature of the video signal makes it very rich in its content, and the subjects in the video have multiple identities depending upon the mode. For example, subjects have a visual identity, i.e. he/she can be identified by face, body language, gait, or clothes captured through the video signal. Subjects also have an audio identity, as a person can be identified by his/her voice. Our scheme tries to hide both visual and audio

identity. Figure 4.2 shows a schematic of the privacy protection scheme. As shown in figure 4.2, under the privacy protection mode of our system the *Face Detection and Blocking* module detects and tracks the subject's face continuously, and blocks it with a solid-color box in real time. We chose not to block out the entire body such as that in [71] as it will essentially remove all visual information in an interview sequence. Face de-identification scheme described in [14] that repaints the face with a generic face is too complicated to be implemented in real-time. To protect the audio identity of the subject, the system identifies the subject's voice by a segmentation algorithm and then distorts it using the PitchScaleSOLA algorithm described in [68]. This part is achieved by *Audio Segmentation* and *Audio Distortion* modules. The distortion is performed in such a way that it conceals the identity of the speaker, but also maintains the intelligibility of the speech, and tries to make different distorted voices sound as much alike as possible to create ambiguity in the distorted data. Thus our scheme protects the privacy of subjects being captured in the life-log video while keeping the recording useful for its review. The details of these algorithms are described in the next two sections.

4.2 Audio Segmentation and Distortion

We propose a simple segmentation algorithm to detect the subject's voice by using the audio signal power. As discussed in the hardware section 3.1 of our life-log system, the microphone in the system is closer to the producer than to the subject. As a result, the audio signal power will generally be higher when the producer is speaking. Therefore, we can separate the audio signal into the segments of the producer speaking



S: Subject (The person who is being recorded)

P: Producer (The person who is the user of the system)

Figure 4.2: Privacy Protection Scheme

and the subject speaking based on thresholding the signal power. Let s_i be the audio sample at time i. We compute the power by first partitioning the audio signal into equal-duration frames of size T and compute its power P_k as follows:

$$P_k = \frac{1}{T} \sum_{i=kT+1}^{kT+T} {s_i}^2 \tag{4.1}$$

where k is the frame index.

The classification is based on two thresholds: a silence-threshold T_S to identify the ambient noise and an producer-threshold T_P to identify the producer's voice. The segmentation scheme is illustrated in Figure 4.3. The two threshold divides the domain of power into three ranges corresponding to silence, subject speaking, and producer speaking. If the power P_k of the audio frame is smaller than or equal to T_S, the frame indicates a pause or a silence in the audio signal. If the power P_k is between T_S and T_P, it is more likely that the subject is speaking in the frame. The corresponding audio frame is processed by *Audio Distortion* module to remove the identity information of the subject in it. If the power exceeds producer threshold T_P,

signifying the producer's voice, the audio frame is not distorted. The frame signifying silence is handled carefully, as it could signify two different things, the frame could be true silence, or it could be a pause in a speech while the producer or the subject is speaking. Thus, we consider the state of the previous frame while deciding whether to distort the frame or not. It also smooths out the segmentation results by reducing false detections of subject speaking when segmentation algorithm confuses the frame as subject speaking because of sudden drop of power due to a pause in the speech of the producer or the subject.

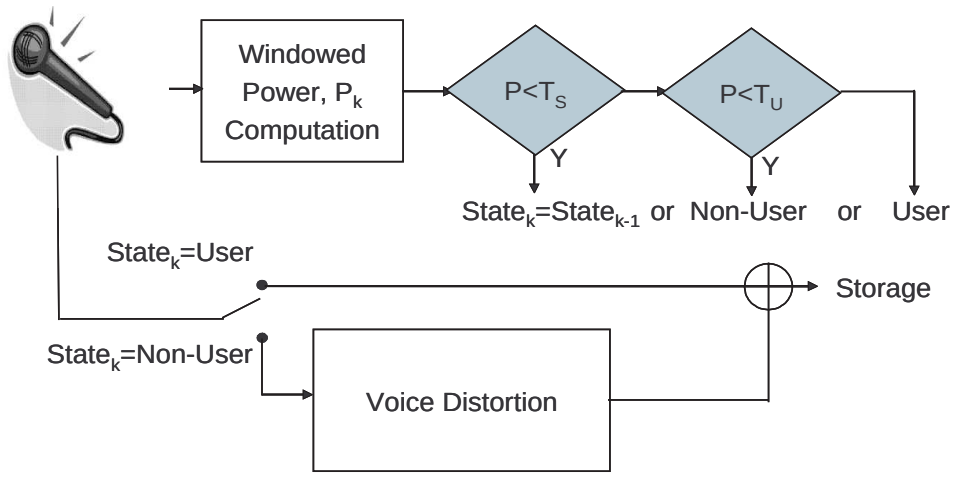


Figure 4.3: Audio Segmentation

The next step is to conceal the identity of the subject by distorting the subject's speech while preserving the intelligibility of the conversation. First we need to define what do we mean by an audio identity of a person. Every person's voice has a fundamental frequency i.e. pitch associated with it, which uniquely identifies the person. We define the audio identity as the pitch of a person's voice. Simple scrambling of audio signal can protect the audio identity of a person perfectly, but it will render the recording useless as one can not understand the conversation. The

dual purpose of protecting the privacy and preserving the intelligibility is achieved by shifting the pitch of the voice signal. We use the time-domain pitch shifting method called PitchScaleSOLA as discussed in [68]. Compared with other pitch shifting algorithms based on frequency domain analysis or delay line modulation, this time-domain method is computationally less complex and thus more amenable to real-time implementation. This algorithm works by first time-stretching the audio signal followed by a re-sampling process to maintain the same length. The time-stretching algorithm expands the input signal from length N_1 to N_2. To preserve the speech structure, the input signal is divided into overlapping blocks of size N with hop size S_a, then the pitch of the overlapping blocks are shifted according to scaling factor $\alpha = N_1/N_2$. α lies between 0.25 and 2 with value 1 signifying no pitch shift. Discrete-time lag of maximum similarity is calculated in the overlapping region. At the point of maximum similarity, the overlapping blocks are weighted by a window function and then summed together. The re-sampling process is performed with an inverse sampling ratio of N_1/N_2 so as to undo the changes in the number of samples. The next section 4.2.1 provides in detail explanation of operation of the pitch shifting algorithm. We noticed that out of various parameters, such as the scaling factor α, block length N, and hop size S_a; the parameter α affects the quality of the distortion and the intelligibility of the distorted speech the most. In the next chapter 5 *"Performance Evaluation of Privacy Protection Module"* sections 5.1.2 and 5.2.1, we discuss experiments to study the effect of using different parameters in the audio distortion algorithm.

4.2.1 Pitch Shifting Algorithm

We decided to use the pitch shifting algorithm called PitchScale SOLA(Synchronous Overlap and Add) discussed in [68] to implement subject voice distortion. PitchScaleSOLA algorithm involves two processes. The process first involves time stretching or time scaling and second process conducts re-sampling, as shown in Figure 4.4. The order of the processes can be interchanged.

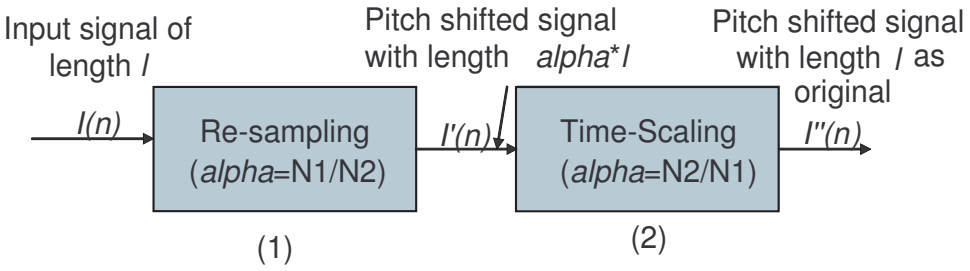


Figure 4.4: Pitch Distortion Algorithm (two processes)

In the first process of *Time stretching*, a time scaling method expands or compresses the input signal from length N_1 to N_2. This process does not change the pitch of the signal. Spectral envelop of the signal remains same. The time stretching of the speech signal is achieved by a simple correlation based method called "Synchronous Overlap and Add(SOLA). Description of the algorithm is as follows:

Step (1): Input signal is divided into equal size windows of block length N with time shift of S_a as shown in figure 4.5.

Step(2): The equal sized blocks are repositioned by shifting it with factor S_s which is equal to $\alpha * S_a$ as shown in figure 4.6, where α is time scaling factor.

Step(3): Cross-correlation between overlapping regions (I_{L1} and I_{L2}) of two blocks ($I_1(n)$ and $I_2(n)$) respectively, is calculated as follows,

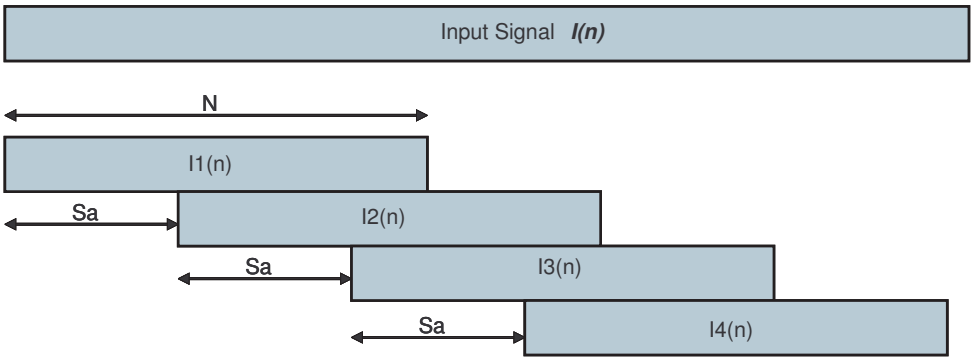


Figure 4.5: Time Stretching algorithm, Step 1

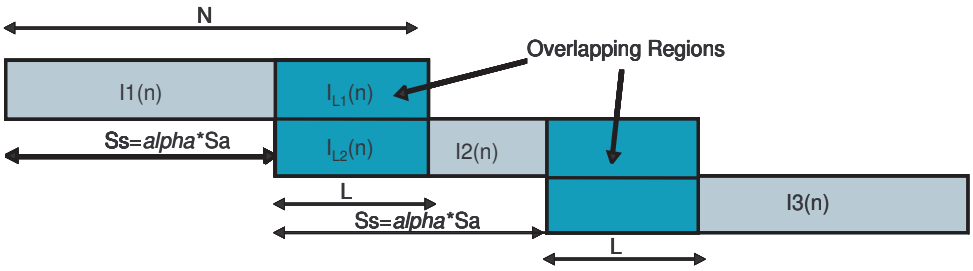


Figure 4.6: Time Stretching algorithm, Step 2

$$r_{I_{L1}I_{L2}}(m) = \frac{1}{L}\sum_{n=0}^{L-m-1} I_{L1}(n).I_{L2}(n+m), 0 \leq m \leq L \tag{4.2}$$

In above equation, L is the length of the overlap interval as shown in figure 4.7.

Step(4): Discrete-time lag k_m is estimated at where the cross-correlation,

$$r_{I_{L1}I_{L2}}(k_m) = r_{\max} \tag{4.3}$$

has its maximum value, as shown in figure 4.8.

Step(5): The overlapping regions are again shifted at the point of maximum cross-correlation lag, as shown in figure 4.9.

Step(6): Fade-in and Fade-out functions are calculated for the new overlapping regions after shifting to the point of maximum correlation, as shown in figure 4.10.

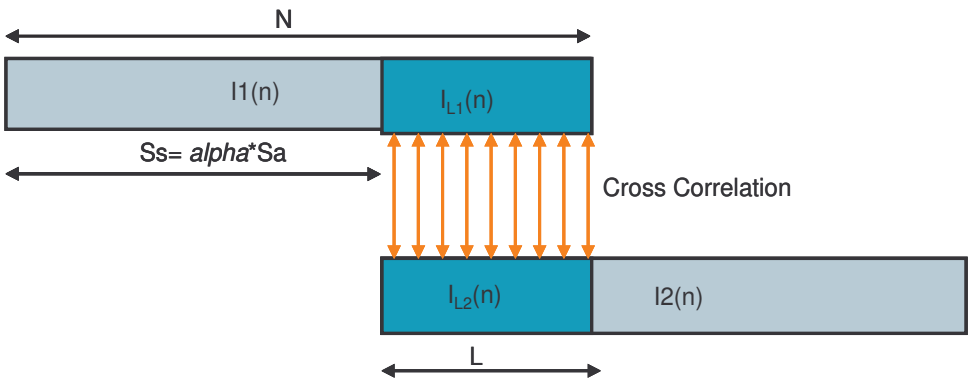


Figure 4.7: Time Stretching algorithm, Step 3

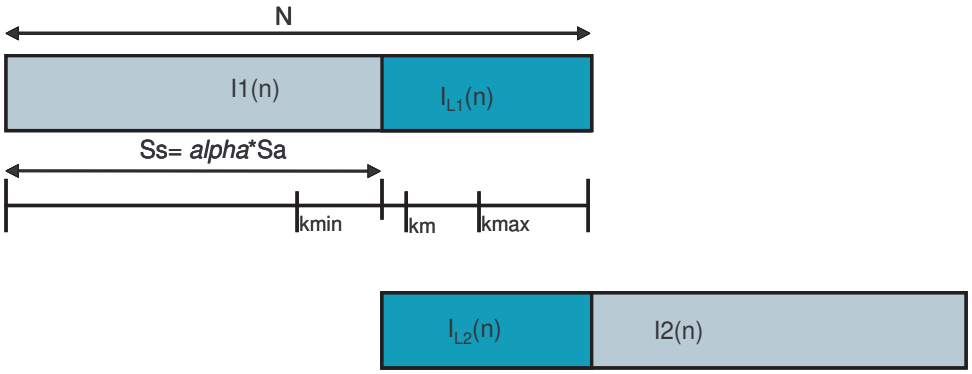


Figure 4.8: Time Stretching algorithm, Step 4

Step(7): The overlapping regions of blocks, (I_{L1} and I_{L2}) are weighted by fade-in and fade-out functions in overlapping regions and finally added to get the final time stretched output signal, as shown in figure 4.11.

The SOLA algorithm is relatively less complex and is based on three parameters S_a, N and α. All of these parameters are independent of the pitch period of the input signal.

In the second process involving *Re-sampling*, the expanded or compressed output audio signal of the first process is re-sampled to make the length of signal from N_2 to the original N_1. Re-sampling of the signal causes its pitch to shift. The spectrum of

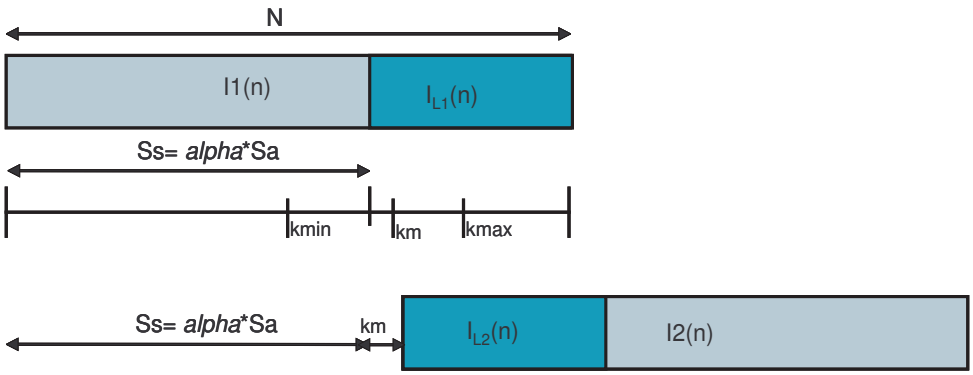


Figure 4.9: Time Stretching algorithm, Step 5

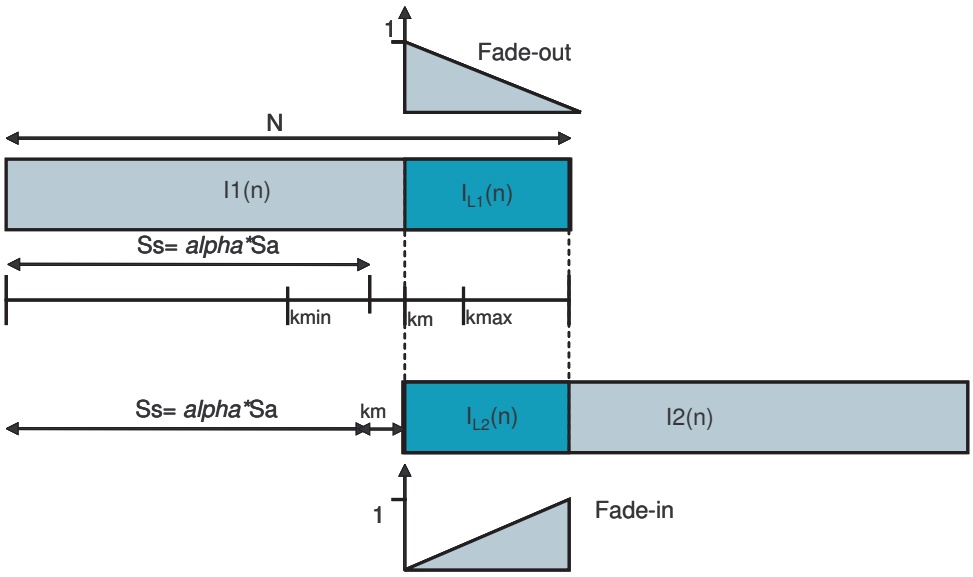


Figure 4.10: Time Stretching algorithm, Step 6

the sound is compressed or expanded over the frequency axis depending on α, which results in pitch-shifting of the signal. Also harmonics are repositioned in the spectra, while keeping relations between them unchanged. The relations between harmonics are only scaled by the time scaling factor α.

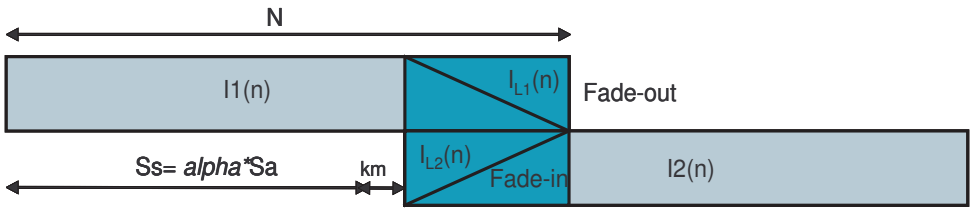


Figure 4.11: Time Stretching algorithm, Step 7

4.3 Face Blocking Module

The figure 4.12 shows the operation of our *Face Detection and Blocking* module. Face detection is based on efficient implementation of the Adaboost face classifier by Viola and Jones [66] in the OpenCV software package [72]. This implementation is very efficient on our micro PC; it is capable of identifying most of the upright frontal faces under good lighting condition at the rate of 15 frames per second for a frame size of 352×288. Applying this classifier on a frame-by-frame basis, however, is not accurate enough for privacy protection. Whenever a person turns his/her head or makes any hand gesture that partially occludes the face, the classifier fails to detect the face. Furthermore, the performance of the classifier is adversely affected by the movement of the camera, which is inevitable as the camera is mounted on the shoulder of the producer. Such momentary relapse is usually sufficient for a viewer to identify the subject. To further improve the performance, we have added a temporal tracking component using the classifier's outputs as observations. The tracking component is based on tracking the skin color measured by the dominant hue color in the face region identified by the classifier. If the classifier fails to provide any observation, we search for all pixels that match the skin color in an area slightly larger than the last-observed face region. The new face region is defined as the bounding box

containing these pixels. If no such bounding box is found, the face is declared to have disappeared. Occasionally, there are background objects that resemble the skin color and the proximity of these objects with a face may introduce false tracks. To limit the lifetime of these false tracks, we mandate that all face tracks must be validated by a true face observation from the classifier within a certain time limit, empirically set to three seconds in our system. If there are multiple face observations in a scene, we match these observations to existing tracks based on minimizing the sum of their distances on the image plane.

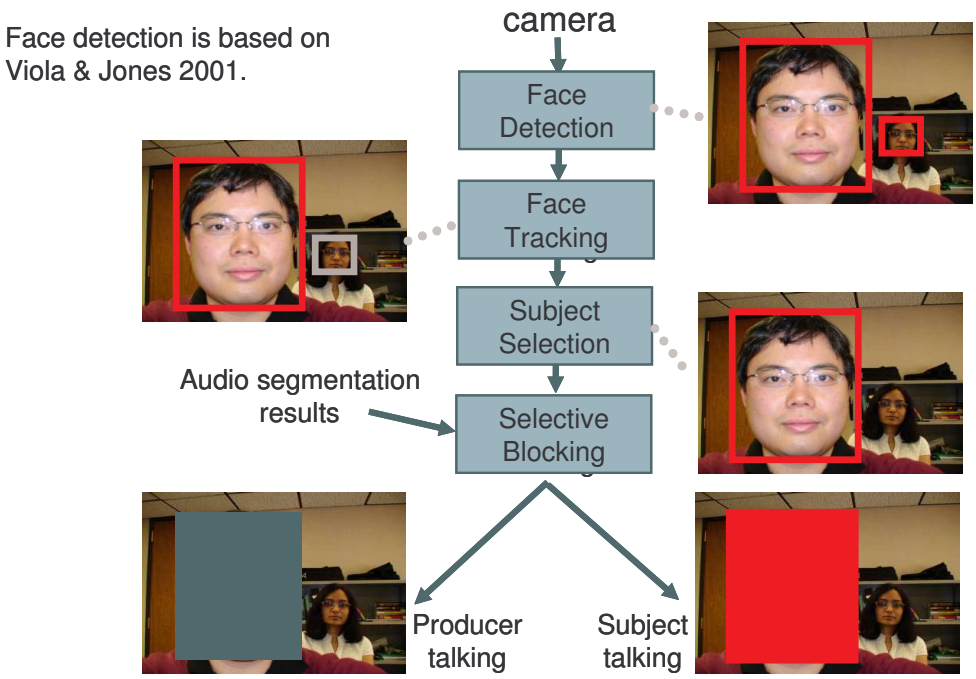


Figure 4.12: Face Detection and Blocking Module

The final step is to obfuscate the identified face region. We chose to color the entire region using a single color as it reveals no information about the underlying face. In order to provide a visual cue on whether the subject is speaking, we utilize the output from the audio segmentation algorithm described in Section 4.2 to change the blocking

color from black to red when the subject starts talking. This simple step provides better visual feedback and indicates that the accompanied voice is being distorted. In the figure 4.12, shows selective blocking of subject faces. But we currently do not perform selective blocking if multiple faces are present. However, this could be done quite easily by having the producer to select the particular subject of interest.

Chapter 5

Performance Evaluations

There are three major components in our privacy protection scheme. The first component is the *audio segmentation algorithm* which plays an important role of detecting the audio segment when the subject is speaking. The second component is the *pitch shifting algorithm* which distorts the subject speaking segment to hide the identity of the subject and along with that it keeps the recording useful and intelligible. The third component is the *face detection and blocking module* which detects and blocks the subject's face to provide privacy protection. The experiments focus on evaluating the performance of audio distortion for privacy protection and usability. Neither the face blocking algorithm nor the pitch shifting algorithm are novel inventions. While there have been evaluation studies on visual privacy protection [16], we are not aware of any studies on audio privacy protection. The goal of this chapter is to provide a detailed study on how well a pitch shifting algorithm can protect individual audio privacy while maintaining the usefulness of the audio signal.

We have conducted three different types of experiments. First, we evaluate the audio segmentation algorithm for its accuracy in dividing the audio signal into subject-voice segments and producer-voice segments. Along with that we also evaluate our face detection and blocking module. Secondly, we evaluate the audio distortion algorithm i.e. pitch shifting algorithm for its ability in protecting the voice identity of a subject. In the third experiment, we evaluate the pitch shifting algorithm for usability of the recording after distortion.

Our experiments show that a good balance between privacy protection and usability can be obtained by implementing our scheme. The pitch-shifting algorithm achieves 100% speaker identification error and thus allow perfect audio privacy protection. To determine if the recording is useful after distortion, we compare intelligibility of the conversation before and after the distortions in terms of Word Error Rate(WER). The average WER before and after distortion (with $\alpha = 1.40$) are 14.2 and 14.4 respectively. Statistical analysis shows that the clarity of the recording for pitch scaling factor $\alpha = 1.40$ is same as that of the original audio clips. Analysis of segmentation algorithm gives recall between 0.86 to 1 in detecting the subject speaking portion of the audio.

In the next section we discuss our initial results. In section 5.2 we describe in detail the more elaborate experiments.

5.1 Initial Experiments

Our early experiments, which were conducted on a smaller scale, showed very good results and have been published in our paper at IEEE SAFE 2007 workshop on the Signal Processing Applications for Public Security and Forensics [73]. For the initial experiments we used two different sets of data. We used our life-log system to capture three interview video sequences with three different subjects in a relatively quiet meeting room. The interview was scripted and the subject and the producer were asked to get familiar with the script before capturing the video. All the interviews had identical contents and each one is of about 1 minute and 30 seconds duration. In Section 5.1.1, we use this data to qualitatively demonstrate the operations of the

entire system, and to quantitatively measure the segmentation algorithm. These video clips are available for download from our group web site `http://www.vis.uky.edu/mialab`. We also collected voice samples of eleven people with two audio clips of each person's voice. This data was used to test the audio distortion algorithm. One of the voice sample of each person is used to train the speaker identification software, and the other one is used for testing purpose. The test results are discussed in Section 5.1.2. The same data set was also used to test the usability of the recording after distortion as discussed in section 5.1.2.

5.1.1 Analysis of Segmentation Algorithm

First we evaluate the performance of the segmentation algorithm discussed in section 4.2. The figure 5.1 shows a plot of the audio signal of one of the three interview sequence that we captured by our life-log system. We use the precision and recall measures to statistically analyze the segmentation output. In Figure 5.1, the transitions are shown with the labels: P $\rightarrow$ Producer Speaking, and S $\rightarrow$ Subject Speaking. It can be seen that the signal energy is significantly greater when the producer is speaking than when the subject is speaking. The precision and recall metrics are computed by counting the number of transitions between producer-voice segments and subject-voice segments in the output and comparing them with those found in the ground truth. The definition of recall and precision are given in Equations (5.1). The ground truth is manually measured from the videos by noting down the time stamps of transitions between producer-voice segment and subject-voice segment.

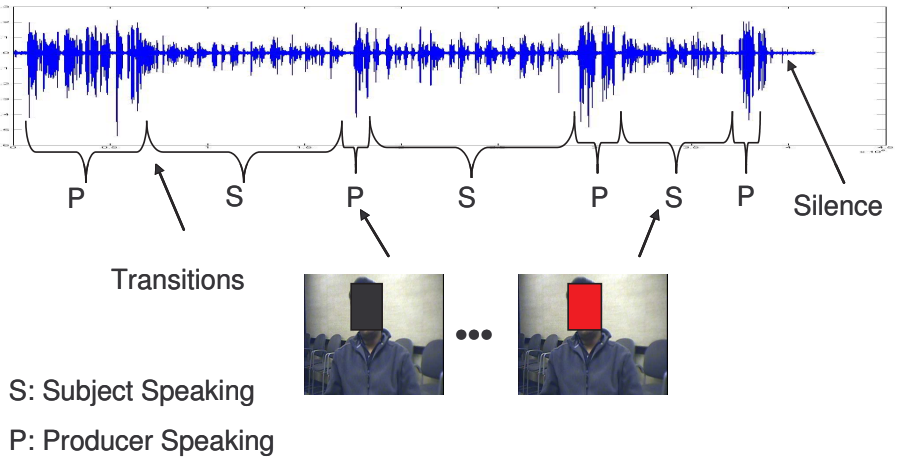


Figure 5.1: Audio Segmentation showing transitions between Subject speaking and Producer Speaking

$$\begin{aligned} \text{Recall} &= \frac{\#\text{ correctly-identified transitions}}{\#\text{ transitions in ground-truth}} \\ \text{Precision} &= \frac{\#\text{ correctly-identified transitions}}{\#\text{ identified transitions}} \end{aligned} \tag{5.1}$$

To compare the performance of the segmentation algorithm, we used CMU audio segmentation method discussed in the paper [74] as a benchmark. The paper presents a system that automatically transcribes a broadcast video and consists of audio segmentation as one of the step in the process. The segmentation is performed by modeling number of speakers; authors call it as acoustic models. Sliding window was used to extract the segmentation points in the audio stream where there are changes in acoustic models. We used a software provided by NIST which is an implementation of this algorithm for our experiments. The results of our segmentation algorithm and CMU segmentation algorithm are shown in Table 5.1.

As shown in Table 5.1, our segmentation algorithm produces good recall performance but rather poor precision values. Our algorithm generates extra transitions

Table 5.1: Results of Segmentation Algorithm

	Our Algo		CMU Algo	
Meeting#	Precision	Recall	Precision	Recall
1	0.375	0.8571	0.667	0.57
2	0.583	1	1	0.57
3	0.353	1	0.4	0.5

because it sometimes confuses pauses in a person's speech as transitions. This problem can potentially be alleviated by adaptively adjusting the size of the window in measuring the signal power. On the other hand, the CMU segmentation algorithm gives better precision, as it has fewer false detections. But CMU algorithm gives poor recall, which might not be good for the privacy protection, as the segment where the subject is speaking might be undetected and can disclose the identity of the subject. In terms of privacy protection, low precision has lower impact than lower recall, thus we argue that our algorithm is better for this application. Again, CMU has greater complexity than our algorithm which is less desirable for real time application like the life-log system.

Our face detection and blocking module works reasonably well. Selected frames are shown in Figure 5.2. In all three sequences, the faces are blocked at all time. There are occasional false alarms that linger for a short period of time. This does not have any adverse effect on privacy protection.

5.1.2 Analysis of Audio Distortion Algorithm

In this section we discuss and analyze the audio distortion algorithm.

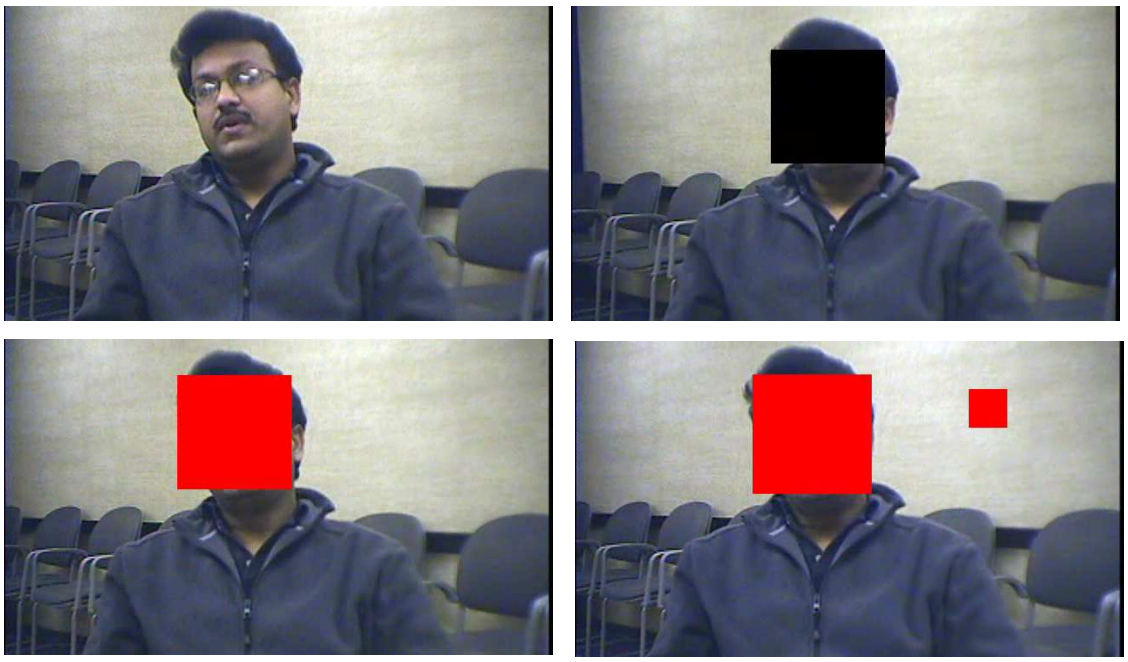


Figure 5.2: Top-left: a frame from the original sequence; Top-right: face blocked when the subject is not speaking; Bottom-left: face blocked when the subject is speaking; Bottom-right: false alarms on the background wall.

Privacy Experiments

We first analyze the performance of the audio distortion algorithm based on how well a speaker can be identified after the distortion. We used a public domain text-independent speaker recognition software [75]. We collected two voice samples from eleven test subjects. The first sample was used to train the speaker identification software, and the other one was used for testing. The results are shown in Table 5.2. We ran the speaker identification program on four sets of data: original test data without distortion and three different sets of distorted data with parameters as indicated in the Table 5.2. These three sets of parameters were chosen with the intention that they would give vastly different distorted sounds. To measure their

performances, we computed the error rate in identifying the correct speaker and the number of distinct speakers found by the speaker identification software.

Table 5.2: Results from Speaker Recognition

Testing (personId)	Ground Truth (PersonId)	Without Distortion	Distortion 1 $N = 2048$ $S_a = 256$ $\alpha = 1.5$	Distortion 2 $N = 2048$ $S_a = 300$ $\alpha = 1.1$	Distortion 3 $N = 1024$ $S_a = 128$ $\alpha = 1.5$
1	1	1	5	8	5
2	2	2	6	8	6
3	3	3	5	3	5
4	4	4	6	6	5
5	5	5	3	10	6
6	6	6	8	6	5
7	7	7	5	2	5
8	8	8	10	11	5
9	9	9	5	8	5
10	10	10	5	2	5
11	11	11	4	8	5
Error Rate		0%	100%	90.9%	100%
# Distinct identities		11	6	6	2

We find that setting α around 1.5 produces good error rate (column 4 and 6) and maintains reasonable intelligibility of the conversation. The 100% error rate shows that the voice-distortion algorithm works well in hiding the identity of the speaker. Thus, we conclude that the parameters $N = 1024$, $S_a = 128$ and $\alpha = 1.5$ produce the best overall privacy protection results in our audio distortion algorithm. The last row of the Table 5.2 shows the number of distinct identities recognized by the speaker identification software. It measures the degree of ambiguity created by pitch shifting algorithm in identities in the data set. Lesser the number of distinct identities better the privacy protection as it increases the ambiguity.

Usability Experiments

In our second experiment on the distortion algorithm, we evaluate the intelligibility of the conversation after the distortion. We have attempted to reproduce the transcripts for both the original and distorted audio by using a speech recognition software. However, we were unable to obtain any reasonably correct transcription on the distorted speech due to its un-natural audio characteristics. As a result, we had to resort to manual transcription. Using two different speech samples from eight test subjects, we used the best distortion algorithm to distort one sample while keeping the other unmodified. Five human testers were asked to transcribe the distorted and non-distorted audio files of the eight subjects. We used the standard measure, called Word Error Rate (WER), which commonly used to evaluate the speech recognition system. In our experiments, WER is used to determine the degree of intelligibility of the speech before and after the distortion. We used a tool called SCLITE which is part of NIST's Speech Recognition Scoring Toolkit (SCTK) [76], to calculate WER of a transcription of an audio clip. In general correctly recognized words in a transcription of an audio clip by a speech recognition software are first aligned with the words in reference/correct sequence of words of the clip, and WER is calculated by the formula in the equation 5.2.

$$WER = \frac{S + D + I}{N} \tag{5.2}$$

In the above equation, S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference transcription.

We calculate WER for each transcription by each tester to measure the intelligibility of words before and after distortion. The average WER for each subject is shown in Figure 5.3. WER of distorted speech range between 3% to 43%, which shows that while the distortion has reduced the clarity of speech, it maintains certain level of intelligibility. For certain subjects (4,5,7,8), the difference in WER between undistorted and distorted voices is small. On the other hand, the difference is large for other subjects (1,2,3). One possible reason for this large difference is that subjects (1,2,3) have strong accents as none of them are native speakers of English. The effect of accents on audio distortion is a subject that deserves further investigation.

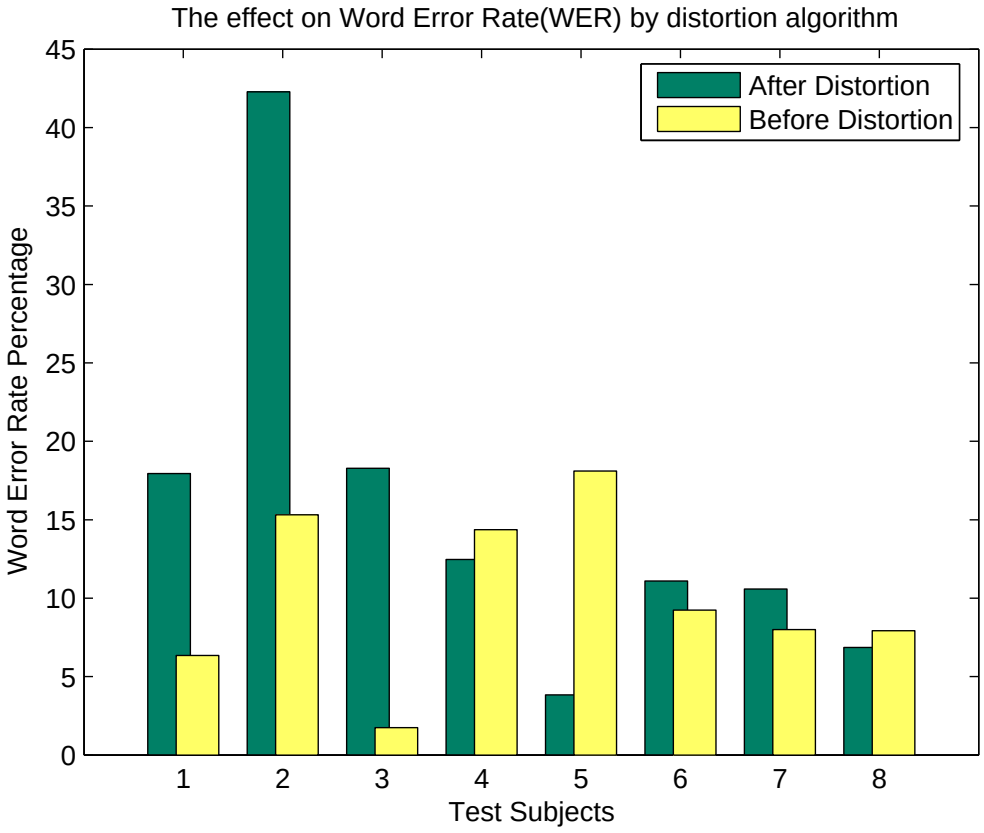


Figure 5.3: The effect of Distortion on Word Error Rate

5.2 Extended Experiments

Our initial small scale experiments showed promising results on the performance of the privacy protection module. To validate the life-log system more robustly, we decided to extend the data set for our next experiments. We conducted different experiments on similar lines as the initial experiments, as discussed in section 5.1 but with larger data set. For our privacy and usability experiments we decided to use well known speech corpora called "(TIMIT) Texas Instruments and Massachusetts Institute of Technology" data set [77]. The TIMIT speech corpora is commonly used for speech recognition and speaker identification experiments by research communities, and consists of speech recordings of 630 speakers taken from eight major dialects regions in the North America. For each speaker there are ten recording with phonetically rich sentences spoken, thus the TIMIT data set has 6300 speech clips in total. The following sections explain the experimental setup and the results.

5.2.1 Subjective Experiments

As explained in section 5.1.2, the usability experiments were conducted by manually transcribing the non-distorted and distorted audio clips by testers (the participants in the experiments). We again resort to manual transcriptions here. The TIMIT data set has well documented transcriptions of all the audio recordings which are required to calculated WER and were used as reference/correct transcriptions. In addition to transcriptions of audio clips, we have expanded the scope of the experiment, and we call it as Subjective Experiments. We sought the tester's help to assess the privacy protection in our system. We asked the testers if they can identify a

person in an audio clip after the distortion. Given a set of distorted or non-distorted audio clips, we asked testers to identify the distinct number of voices in that set. The details of these experiments are given in the next section *Experimental Setup*.

Experimental Setup

As explained in section, 4.2.1, the pitch shifting ratio *alpha*, determines the degree of distortion of the audio signal, i.e. the amount of the shift in the pitch. The value of *alpha* varies between the allowable range 0.2 to 2, and $alpha = 1$ signifies no pitch shift. In order to decide which *alpha* is required to keep the balance between usability of recording and privacy protection capability, we decided to experiment with five different *alpha* values distributed evenly in the allowable range of *alpha*. The values are as follows; $alpha = 1$ (Original voice without any distortion), and $alpha = 0.5$, $alpha = 0.75$, $alpha = 1.25$, and $alpha = 1.40$. For our experiments, we chose 30 speakers from TIMIT data set taken evenly from all eight major dialect regions with comparable number of male and female speakers from each region. We considered five speech samples for each speaker and paired one speech sample with one *alpha* value. For each *alpha* parameter, we created a distorted data set of 30 speech recordings by taking one speech sample of a speaker out of his/her five speech samples. Therefore, we have five sets of speech samples corresponding to five *alpha* values, and each set consisting of speech samples of the same 30 individuals, thus total 30x5 = 150 speech samples. We name the collection of speech samples distorted with pitch scaling parameter $alpha = 1$, $alpha = 0.5$, $alpha = 0.75$, $alpha = 1.25$,

and *alpha* = 1.40 as set A, set B, set C, set D, and set E respectively as shown in Table 5.3.

Table 5.3: Sets and their associated alpha values

Set	Alpha value
A	1
B	0.5
C	0.75
D	1.25
E	1.40

The audio clips from all the data-sets i.e. set A, B, C, D and E, are then divided into five different groups each containing randomly selected six audio clips from each set, such that each group has 30 audio clips. Thus each group consist of five subsets one for each *alpha* value. Each group was assigned to three testers to analyze it. We have total 15 testers. All the experimental data and setup details can be found on our web site `http://vis.uky.edu/~jayashri/transcription_experiments.htm`. Each tester was asked to perform three different tasks.

1. Task 1: The tester was asked to transcribe all audio clips present in the assigned group.

2. Task 2: The tester was asked to identify the number of distinct voices in each subset included in the assigned group.

3. Task 3: For each clip from subset of Set A (which is the original un-distorted speech set); the tester was asked to identify a clip in other subsets in which the same speaker may be speaking.

The purpose of task 1 is to analyze the pitch shifting algorithm for its ability in preserving clarity of a speech after the distortion. The usability experiments discussed

in section 5.1.2 are extended by implementing this task. In this experiment, we analyze how the degree of distortion varies with the *alpha* value and ultimately affects the clarity of the speech in the audio. Next subsection explains the experimental results for task 1. The task 2 and 3 results are discussed in section 5.2.1 and 5.2.1 respectively.

Results from Usability Experiments (Task 1)

We used WER as a measure of intelligibility in the usability experiments similar to as discussed in section 5.1.2. The transcriptions collected from task 1 of subjective experiments 5.2.1 are used to calculate WER for each audio clip. We used a tool called NIST's SCLITE [76] software, to calculate WER. In each group, every clip has three WER associated with it corresponding to transcriptions by three different testers. We calculate the average WER as effective WER for each clip. The Figure 1 explains this experimental step for the group 1. Similarly for groups 2, 3, 4 and 5, results are as shown in Figures 2, 3, 4. and 5 respectively. All the tables are included in Appendix A.

We combine the results from all groups and write down an average WER for audio clips of each person. The Table 5.4 shows the results for each person and for all the sets. The calculated average WER is associated with the corresponding person ID for the respective set. The average WER for each set is shown in the last row of Table, and is obtained by averaging the WER of each person for that *alpha* value. The bar chart in Figure 5.4 shows the average WER for each person corresponding to different *alpha* values.

Table 5.4: Average WER for each Set

Person Ids	Alpha 1 (Set A) ($\alpha = 1$)	Alpha 2 (Set B) ($\alpha = 0.5$)	Alpha 3 (Set C) ($\alpha = 0.75$)	Alpha 4 (Set D) ($\alpha = 1.25$)	Alpha 5 (Set E) ($\alpha = 1.4$)
1	23.83	100.00	18.90	23.83	28.23
2	17.80	100.00	36.13	13.33	16.20
3	15.27	100.00	14.53	22.23	18.50
4	9.53	100.00	0.00	9.53	3.70
5	0.00	100.00	9.53	9.53	14.30
6	16.67	100.00	23.10	21.23	7.40
7	8.33	100.00	0.00	20.53	6.37
8	28.60	100.00	20.00	30.50	33.33
9	8.03	100.00	22.23	25.73	0.00
10	12.23	100.00	15.17	26.30	18.53
11	22.23	100.00	4.77	0.00	25.00
12	20.83	100.00	26.67	22.23	6.67
13	19.70	100.00	33.37	25.00	16.20
14	14.30	100.00	7.70	0.00	16.67
15	14.30	100.00	19.17	11.13	22.23
16	11.63	100.00	50.00	22.23	0.00
17	10.43	100.00	52.77	5.57	0.00
18	19.47	100.00	25.00	20.63	5.40
19	2.23	100.00	38.13	3.33	9.53
20	9.37	100.00	27.77	16.70	16.67
21	7.40	100.00	50.00	14.07	14.30
22	7.13	100.00	20.00	0.00	10.00
23	24.57	100.00	44.43	14.30	22.20
24	13.07	100.00	12.73	9.10	16.67
25	15.17	100.00	13.33	17.83	9.10
26	11.13	100.00	14.13	33.33	14.30
27	21.43	100.00	5.57	6.67	6.67
28	25.00	100.00	20.00	6.67	33.33
29	4.17	100.00	19.03	0.00	22.23
30	11.10	100.00	27.77	26.67	16.67
Avg WER	14.17	100	22.398	15.27	14.347

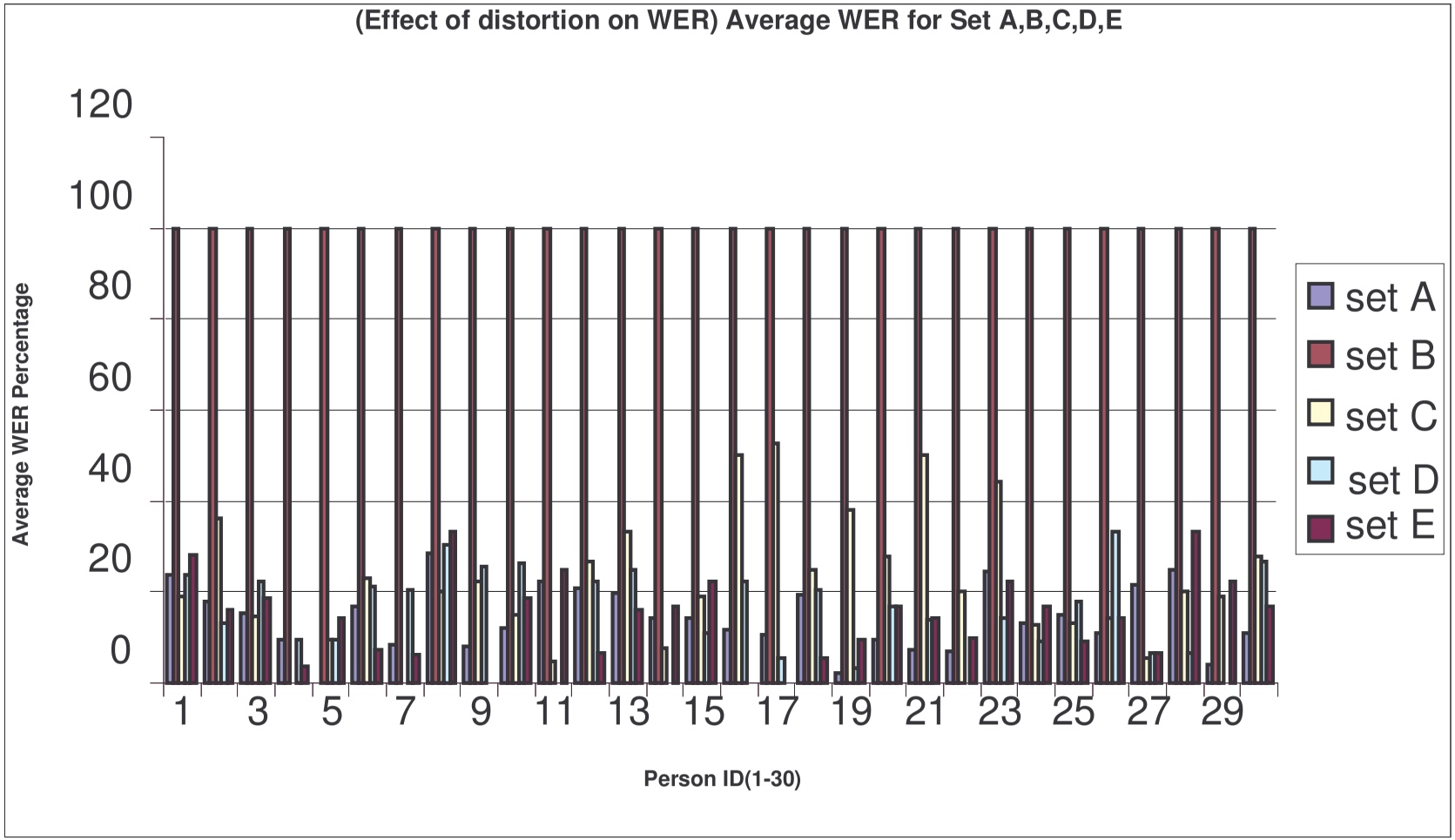


Figure 5.4: Bar chart showing effect of distortion on Word Error Rate (WER)

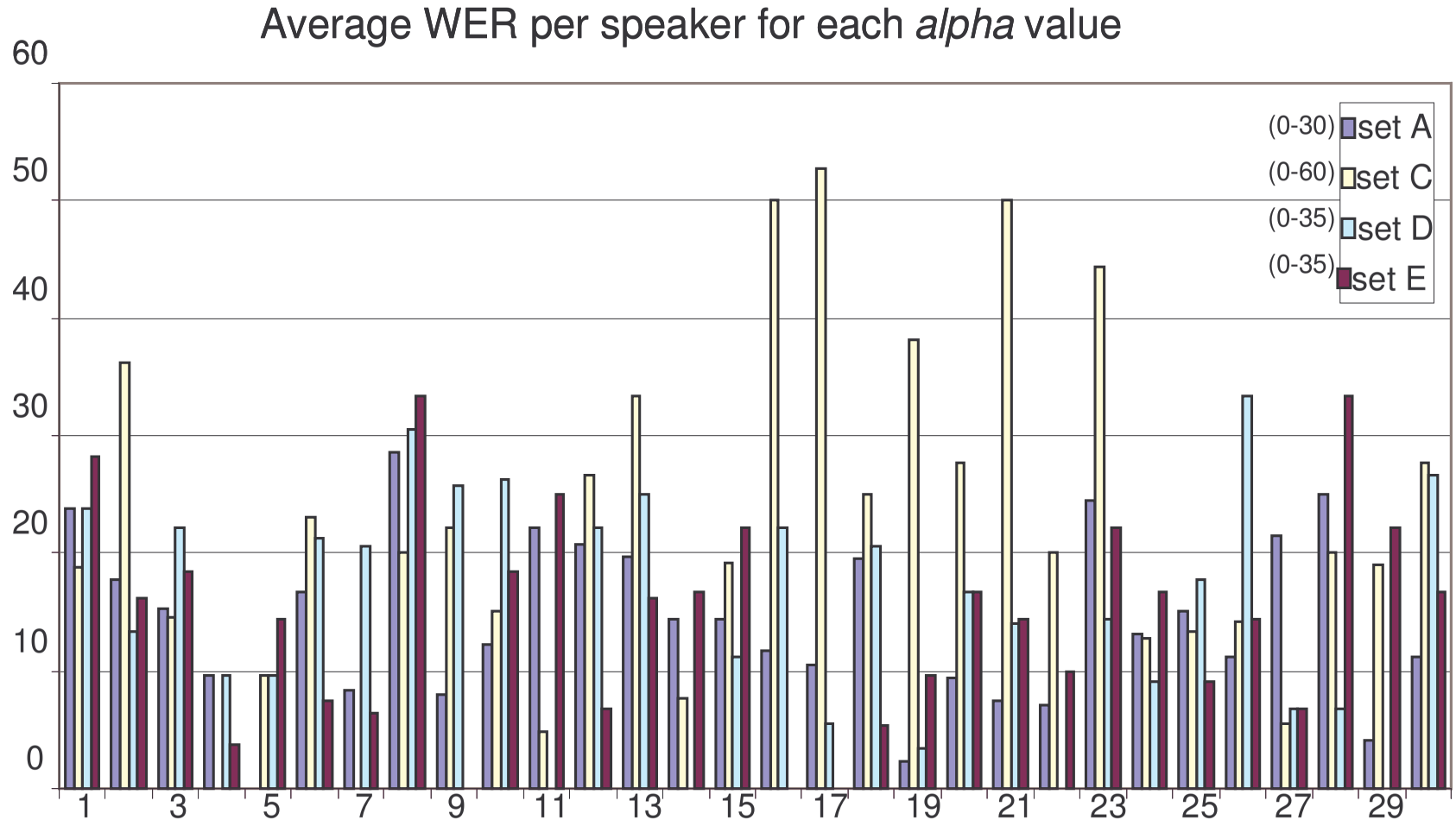


Figure 5.5: Bar chart showing effect of distortion on Word Error Rate (WER) (With Set B removed)

From the bar chart in Figure 5.4 and Table 5.4, it can be seen that the $alpha = 0.5$ (Set B) distorts the speech extremely high as almost all the testers were not able to decipher any of the words spoken in the distorted clips. Thus it gives the worst WER i.e. 100%. The Figure 5.5 shows the same bar chart as in Figure 5.4 but the set B is removed. It shows better comparative effects on WER among the other sets caused by the distortion. It can be seen that the average WER for $alpha = 1$, $alpha = 1.25$ $alpha = 1.40$ ranges between 0% to 30-35%. For $alpha = 0.75$ it ranges from 0% to 60%. The bar chart figure 5.6 which plots average WER for each set A,B,C,D and E, shows that the average WER for set A and E does not change significantly. Thus, we conclude that Set E which corresponds to $alpha = 1.40$ has minimal impact on usability of recording, and has almost same intelligibility as the undistorted speech (Set A). Hence we chose $alpha = 1.4$ for the pitch shifting algorithm in our system.

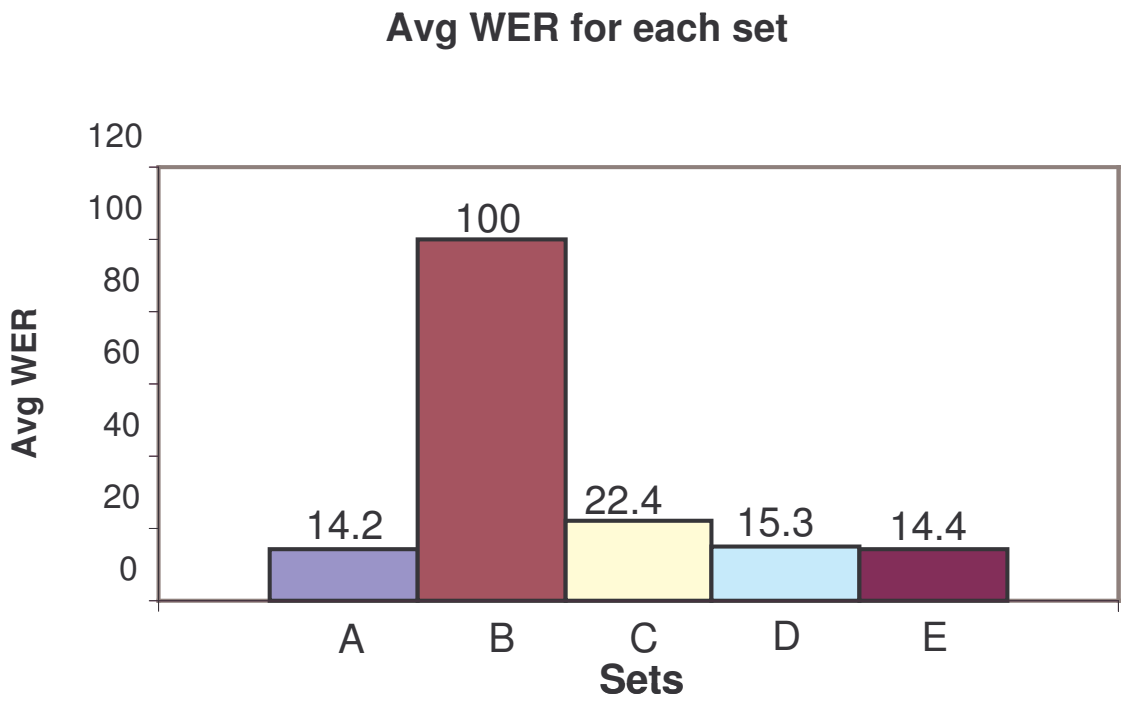


Figure 5.6: Average WER For Set A,B,C,D,E

Statistical Analysis of the data by z-test:

As WER is a ratio, we conduct a z-test based on two population proportions to validate the above results. We want to ascertain that, the differences in the average WER shown in Figure 5.6 are valid, even after taking into consideration the standard deviation in average WER for each set. Our null hypothesis is that the average WER does not change (from Set A) after the distortion for a given value of α level, as follows. Not to confuse with pitch shifting parameter *alpha*; here α level is the probability of making the type I error in z-test. The type I error occurs when we reject the null hypothesis H_0 when it is actually true [78].

$$H_0 : p_1 - p_2 = 0$$
$$H_a : p_1 - p_2 \neq 0$$

Here p_1 is population 1, and p_2 is population 2. Population 1 (p1) is WER of audio clips without distortion (Set A), while Population 2 (p2) is a set of WER of corresponding audio clips after distortion (Set B, C, D, E). As we are considering the difference in two populations in either direction, we calculate two tail z-test estimates.

Each set has 30 audio clips and on an average each audio clip has 12 words in it. Thus population size for two population p1 and p2 is (12*30=360). The mean WER for each set is as shown in last row of the table 5.4. Other parameters such as α, or confidence level, needed to calculate the z-test are as shown in table 5.5. The z-statistics is calculated by standard formula as explained in the book [78]. If the calculated z value for two populations is less than the critical z value (shown in the fourth row in Table 5.5), we say that our null hypothesis H_0 is true. The critical z

value is estimated from z table for confidence level 95%. The rule for rejecting the null hypothesis is shown in last row of the table 5.5.

Table 5.5: Parameters for z-test

Name	Value
Population Size	360
alpha (α)	0.05
Confidence level	95%
z-Test critical	1.96 (from z-table)
Rule for Rejection of H_0	$z \geqslant z_{\alpha/2}$ or $z \leqslant -z_{\alpha/2}$

Table 5.6: Statistical Analysis with z-test, $\alpha = 0.05$

Comparison	z-Test statistics
Set A and B	$46.705 \geqslant 1.96$
Set A and C	$2.873 \geqslant 1.96$
Set A and D	$0.419 \leqslant 1.96$
Set A and E	$0.0695 \leqslant 1.96$

The Table 5.6 shows the z-test statistics for comparison of set A and all the other sets. It is seen that the calculated z value for set D and E (with set A) is less that the critical z value. Thus, the hypothesis H_0 i.e. WER does not change after distortion is true for Set D and Set E. On the other hand for set B and C, the hypothesis does not hold. This again shows that the pitch shifting parameters $alpha = 1.40$ and $alpha = 1.25$ are good parameters to keep balance of usability and privacy.

Task 2 Experimental results

The purpose of Task 2, is to measure an ambiguity created by the audio distortion algorithm. We asked all testers to identify number of distinct voices from each subsets (made of set A, B, C, D and E) belonging to the corresponding assigned group. As explained before in section 5.2.1, each group has 6 randomly selected audio clips from each set A, B, C, D, and E. The average number of distinct voices recognized by testers is shown in last but one row of in Table 5.7. The last row of the Table lists the ambiguity corresponding to each *alpha* value. It is calculated by following formula,

$$A = 1 - m/T \tag{5.3}$$

Here, A is ambiguity, m is average number of distinct voices recognized in the set, and T is actual total number of distinct voices in the set. The value of A equal to 0 signifies no ambiguity and increasing values of A denote increasing ambiguity.

Table 5.7: Task 2 Results (Average number of distinct voices recognized per subset in each group)

	Average # of distinct voices per subset (Each subset consist of 6 audio clips)				
Group#	subset A	subset B	subset C	subset D	subset E
Group 1	6.00	3.33	4.33	4.00	3.33
Group 2	6.00	3.00	3.33	4.00	4.00
Group 3	6.00	2.00	4.00	3.00	4.00
Group 4	6.00	2.67	4.00	3.67	2.67
Group 5	6.00	3.00	3.00	3.67	4.00
Average # (m)	6.00	2.75	3.92	3.67	3.50
Ambiguity(A)	0	0.54	0.34	0.38	0.41

It can be seen from the results, for set A (without distortion) the distinct recognized voices are 6 out of 6. That means it has the least ambiguity. Set B, which

distorts the audio signal with $\alpha = 0.5$ has the most ambiguity (Average number of distinct voices are 2.75 out of 6), which means that most of the voices sound similar after distortion. But as explained in the usability experiments, it does not keep the clarity of the conversation intact. Set D and E have almost same number of recognized distinct voices 3.46 and 3.50 respectively. This again proves, that Set E gives reasonable ambiguity, and is suitable to create anonymous distorted data set with robust privacy protection.

Task 3 Experimental results

The task 3 is a subjective test for privacy protection experiments presented in section 5.1.2. For each clip from set A, we asked testers to identify one clip from subsets B, C, D, and E, in which they think that the same person is speaking. The results from testers showed that none of the speakers from set A was identified from other distorted sets by testers. This again shows that our audio distortion gives 100% recognition error rate in subjective privacy protection experiments. Thus, the pitch shifting algorithm chosen for the privacy protection scheme in our system works well in hiding the identity of the speaker.

Chapter 6

Conclusions

In this work, we presented a practical wearable system, called *"Life-log system"* with an in-built privacy protection scheme. The proposed life-log system is designed for an interview scenario in which the producer is interviewing a single subject. Many design issues, which would be important from a user's point of view such as ease of use, light-weight, etc were taken into consideration in designing the hardware of the system. A simple privacy protection scheme is presented that protects the identity of subjects being recorded in the life-log videos. In this work, we do not claim to provide a full fledged system, but instead a novel framework with initial steps to build user friendly privacy protection mechanism in life-log video is presented.

The major contributions of this work include a privacy protection scheme that implements real time face tracking and blocking mechanism, as well as real time audio distortion of a subject's voice. This scheme was rigorously tested for its feature detection and blocking abilities. We analyze the audio distortion algorithm for its ability to hide the identity of a subject while keeping the speech/conversation clear enough to be useful. Many experiments such as privacy experiments, intelligibility experiments, and subjective experiments were conducted to this end. Our privacy and intelligibility experiments show that the pitch shifting algorithm distorts the speech sufficient enough to hide the identity of a person but retains enough clarity of the speech to keep the recording useful. We also analyze the audio segmentation

algorithm for its accuracy in detecting the subject's voice by precision and recall metrics.

In the future work, this system could be further developed in many ways. The audio segmentation algorithm, which is currently based solely on the power feature could be improved in significant ways. More robust audio features such mel-cepstral coefficients can be used to identify the subject's voice in audio segmentation algorithm. Probabilistic models such as the Hidden Markov Model (HMM) could be used to smooth out the false detections during segmentation. One of the disadvantages of the current audio distortion algorithm in our system is that it is not a reversible process. Therefore, the original recording can never be retrieved. An improvement can be done by making the audio distortion reversible, and by providing a security mechanism so that only an authorized person has access to information needed to extract the original recording. To further improve visual identity protection of a person, we would like to block the whole body by creating a silhouette effect instead of just face blocking, which will keep body language also intact. In summary, this prototype system has great potential and could be deployed into a product which will be very useful to law enforcement, military, and other security and privacy protection applications.

Bibliography

[1] V. Bush. As we may think. *The Atlantic Monthly*, 176:101–108, 1945.

[2] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. Mylifebits: Fulfilling the memex vision. In *Proceedings of ACM Multimedia*, pages 235–238, 2002.

[3] Cylon systems. http://www.cylonsystems.com/.

[4] IBM Global Services. Mobile digital video capture and management: Increasingly indispensable, January 2004.

[5] DARPA(Defence Advanced Research Program Agency. Assist (advanced soldier sensor information system and technology). http://www.ukcrc.org.uk/gcresearch.pdf.

[6] Daniel Solove. *The Digital Person: Technology and Privacy in the Information Age*. NYU Press; New Ed edition, September 2006.

[7] L. Sweeney. Navigating computer science research through waves of privacy concerns: Discussions among computer scientists at carnegie mellon university. In *ACM Computers and Society*, volume 34, April 2004.

[8] Jay Stanley and Barry Steinhardt. Drawing a blank: The failure of facial recognition technology in tampa, florida. AN ACLU(American Civil Liberties Union) SPECIAL REPORT, January 2002.

[9] Dibya Sarkar. Florida city moves to ban face-recognition system. USA Today, 23 August 2001.

[10] Noah Shachtman. Pentagon kills lifelog project. Wired News, February 2004.

[11] William Cheng, Leana Golubchik, and David Kay. Total recall: Are privacy changes inevitable? In *Proceedings of the 1st ACM workshop on Continuous Archival and Retrieval of Personal Experiences (CAPRPE '04)*, pages 86–92, New York, New York, USA, 2004. ACM Press.

[12] J. Wickramasuriya, M. Datt, S. Mehrotra, and N. Venkatasubramanian. Privacy protecting data collection in media spaces. In *ACM International Conference on Multimedia*, pages 48–55, New York, NY, Oct 2004.

[13] W. Zhang, S.-C. Cheung, and M. Chen. Hiding privacy information in video surveillance system. In *Proceedings of the 12th IEEE International Conference on Image Processing*, Genova, Italy, September 2005.

[14] E. N. Newton, L. Sweeney, and B. Main. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, February 2005.

[15] L. Sweeney. k-anonymity: A model for protecting privacy. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, pages 557–570, 2002.

[16] Q. Zhao and J. Stasko. Evaluating image filetring based technique in media space applications. In *In ACM Conference on Computer Supported Cooperative Work*, pages 11–18, Seattle, MA, Nov 1998.

[17] Daniel Ashbrook, Kent Lyons, and James Clawson. Capturing experiences anytime, anywhere. *IEEE Pervasive Computing*, 5(2):8–11, April-June 2006.

[18] Jim Gemmell, Lyndsay Williams, Ken Wood, Roger Lueder, and Gordon Bell. Passive capture and ensuing issues for personal lifetime store. In *Proceedings of The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CAPRPE '04)*, pages 48–55, New York, NY, USA, October 2004.

[19] Steve Mann. Continuous lifelong capture of personal experience with eyetap. In *ACM Multimedia*, New York, NY, October 2004.

[20] B.N. Schilit, N. Adams, R. Gold, M. Tso, and R. Want. The parclab mobile conputing system. In *Proceedings of the Fourth Workshop on Workstation Operating Systems(WWOS-IV)*, pages 34–39, 1993.

[21] M. Lamming and M. Flynn. Forget-me-not: intimate computing in human memory. In *FRIEND21,International Symposium Next Generation Human Interface*, February 1994.

[22] B. Clarkson. *Life Patterns: Structure from Wearable Sensors*. PhD thesis, Massachusetts Institute of Technology, 2002.

[23] Jennifer H. and Rosallind W. Startlecam: A cybernetic wearable camera. In *Proceeding of the Second International Symposium on Wearable Computers (ISWC)*, pages 125–128, 1998.

[24] D. Tancharoen, T. Yamasaki, and K. Aizawa. Practical experience recording and indexing of life log video. In *Capture, Archival and Retrieval of Personal Experiences (CARPE)*, Singapore, 2005.

[25] ACM Workshop. Continuous archival, retrieval of personal experience. http://www.sigmm.org/Members/jgemmell/CARPE.

[26] Engineering and Physical Sciences Research Council's (EPSRC). Memories for life: "managing information over a human lifetime". http://www.memoriesforlife.org/.

[27] Edited by Tony Hoare and Robin Milner. Grand challenge in computing research. http://www.ukcrc.org.uk/gcresearch.pdf, 2006.

[28] Aizawa K. and Ishijima K. Summarizing wearable video. In *International Conference of ICIP*, volume 3, pages 398–401, October 2001.

[29] T. Hori and K. Aizawa. Capturing life log and retrieval based on context. In *IEEE ICME*, June 2004.

[30] K. Aizawa, D. Tancharoen, S. Kawasaki, and T. Yamasaki. Efficient retrieval of life log based on context and content. In *Capture, Archival and Retrieval of Personal Experiences (CARPE)*, New York, NY, 2004.

[31] Wei-Hao Lin and Alexander Hauptmann. Structuring continuous video recordings of everyday life using time-constrained clustering. In *IS and T/SPIE Symposium on Electronic Imaging*, San Jose, CA, January 15-19 2006.

[32] D.P.W. Ellis and K. Lee. Minimal-impact audio-based personal archives. In *Capture, Archival and Retrieval of Personal Experiences (CARPE)*, New York, NY, 2004.

[33] Andrew Y. Ng, Micheal Jordon, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of Neural Information Processing Systems*, 2001.

[34] Daniel Ellis and Keansub Lee. Features for segmenting and classifying long-duration recording of "personal" audio. In *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 39 – 47, New York, New York, USA, 2004.

[35] B. Clarkson, K. Mase, and A. Pentland. A recognizing user context via wearable sensors. In *Proceedings of ISWC*, pages 69–75, October 2000.

[36] Antonio Torralba, Kevin Murphy, Willian Freeman, and Mark Rubin. Context-based vision system for place and object recogntion. In *In proceedings of International Conference of Computer Vision*, 2003.

[37] L. Willenborg and T. de Waal. Elements of statistical disclosure control. In *Springer Verlag*, 2000.

[38] J. Chawala, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases. In *Theory of Cryptography Conference*, 2004.

[39] Jelke G. Bethlehem, Wouter J. Keller, and Jeroen Pannekoek. Disclosure control of microdata. In *Journal of the American Statistical Association*, volume 85, pages 38–45, 1990.

[40] R. Brand. Microdata protection through noise addition. In *Inference Control in Statistical Databases, Newyork-Springer*, pages 97–116, 2002.

[41] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *International Conference on Database Theory*, 2005.

[42] L. Sweeney. *Information Explosion. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Urban Institute, Washington, DC, 2001.

[43] Dominic Hughes and Vitaly Shmatikov. Information hiding, anonymity and privacy: A modular approach. In *Journal of Computer Security*, 2004.

[44] Joseph Y. Halpern and Kevin R. O'Neill. Anonymity and information hiding in multiagent systems. In *Journal of Computer Security*, volume 13 of *3*, pages 483–514, May 2005.

[45] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *In Proceedings of the ACM Symposeum on Principles of Database Systems*, pages 202–210, 2003.

[46] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. In *Internation Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, pages 571–588, 2002.

[47] Arik Friedman, Ran Wolff, and Assaf Schuster. Proividing k-anonymity in data mining. Accepted for The VLDB Journal.

[48] M. Boyle, C. Edwards, and S. Greenberg. The effect of filtered video on awareness and privacy. In *In ACM Conference on Computer Supported Cooperative Work*, pages 1–10, Philadelphia, PA, Dec 2000.

[49] F. Dufaux and T. Ebrahimi. Scrambling for video surveillance with privacy. In *In IEEE Workshop on Privacy Research in Vision*, 2006.

[50] F. Dufaux, M. Quaret, Y. Abdeljaued, A. Navarro, F. Vergnenegre, and T. Ebrahimi. Privacy enabling technology for video surveillance. In *In Proceedings of SPIE 6250*, 2006.

[51] C. Neustaedter, S. Greenberg, and M. Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. In *ACM Transactions on Computer Human Interactions on Knowledge and Data Engineering*, volume 17, pages 232–243, 2005.

[52] Scott E. Hudson and Ian Smith. Techniques for addressing fundamental privacy and disruption tradeoffs in awareness support systems. In *Proceedings ACM CSCW*, pages 248–257, 1996.

[53] D.A. Fidaleo, H.-A. Nguyeb, and M. Trivedi. The networked sensor tapestry (nest): A privacy enhanced software architecture for interactive analysis of data in vide-sensor networks. In *In Proceedings of the ACM 2nd International Workshop on Video Surveillance and Sensor Networks*, 2004.

[54] I. Martínez-Ponte, X. Desurmont, J.Meessen, and J.-F. Delaigle. Robust human face hidding ensuring privacy. In *In Workshop on Integration of Knowledge, Semantics and Digital Media Tecnhonolgy*, 2005.

[55] A. Senior, S. Pankati, A. Hampapur, L. Brown, Y.-L. Tian, and A. Ekin. Blinkering surveillance: Enabling video privacy throught computer vision. In *IEEE Security and Privacy*, May/June 2005.

[56] William Luh, Deepa Kundur, and Takis Zourntos. A novel distributed privacy paradigm for visual sensor networks based on sharing dynamical systems. *EURASIP journal on Advances in Signal Processing*, 2007:17, April 2006.

[57] R.S. Fish, R.W. Kraut, R.E.Root, and Rice R.E. Video as a technology for informal communication. In *Communications of the ACM*, volume 36, pages 48–61, 1992.

[58] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pages 655–658, 1988.

[59] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. In *IEEE Transaction on Speech and Audio Processing*, volume 6, pages 131–142, March 1998.

[60] L.C. Schwardt and J.A. Du Preez. Voice conversion based on static speaker characteristics. In *Communications and Signal Processing, 1998. COMSIG '98. Proceedings of the 1998 South African Symposium on*, pages 57–62, September 1998.

[61] John Puterbaugh. http://www.music.princeton.edu/ john/voiceconversion.htm.

[62] Gina Upperman. Performing voice conversion with signal processing. http://cnx.org/content/m12479/latest/.

[63] D. Matrouf, J.-F. Bonastre, and C. Fredouille. Effect of speech transformation on impostor acceptance. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, page 1, Toulouse, May 2006.

[64] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet. Voice forgery using alisp: Indexation in a client memory. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 17–20, March 2005.

[65] Steve Mann. "wearcam"(the wearable camera):personal imaging system for long-term use in wearable computer mediated reality and personal photo/video graphic memory prosthesis. In *International Semantic Web Conference*, pages 124–131, 1998.

[66] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[67] B. Clarkson, N. Sawhney, and A. Pentland. Auditory context awareness via wearable computing. In *Proceedings Perceptual User Interfaces Workshops*, 1995.

[68] Udo Zölzer et al. *DAFX - Digital Audio Effects*. John Wiley and Sons, LTD, 2002.

[69] R. Xiao, M.-J. Li, and H.-J.Zhang. Robust multipose face detection in images. In *IEEE Transaction on CSVT*, volume 14, pages 31–41, January 2004.

[70] Cha Zhang, Yong Rui, and Li-Wei He. Light weight background blurring for video conferencing applications. In *International Conference on Image Processing*, Atlanata, GA, October 2006.

[71] S.-C. Cheung, J. Zhao, and M. V. Venkatesh. Efficient object-based video inpainting. In *Proceedings of the 13th IEEE International Conference on Image Processing*, Atlanta, GA, September 2006.

[72] Open source computer vision library (opencv). http://www.intel.com/technology/computing/opencv/.

[73] J. Chaudhari, S.-C. Cheung, and M. V. Venkatesh. Privacy protection for life-log video. In *IEEE Signal Processing Society SAFE 2007: Workshop on Signal Processing Applications for Public Security and Forensics*, 2007.

[74] Matthew A. Seigler, Uday Jain, Bhiksha Raj, and Richard M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proceedings of the Ninth Spoken Language Systems Technology Workshop*, Harriman, New York, 1997.

[75] Luigi Rosa. Text-independent speaker recognition based on neural networks. http://www.advancedsourcecode.com/neuralnetspeaker.asp.

[76] Nist sclite: Speech recognition scoring toolkit (sctk). http://computing.ee.ethz.ch/sepp/sctk-1.2c-be/sclite.htm or http://www.nist.gov/speech/tools/index.htm.

[77] Timit acoustic-phonetic continuous speech corpus. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1.

[78] Jay L. Devore. *Probability and Statistics for Engineering and the Science*. Brooks/Cole Publishing Company, third edition, 1991.

Appendix A

The experimental data for usability experiments:

For Group 1

Set A	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
seta_1	28.6	28.6	14.3	23.83333333
seta_2	20	26.7	6.7	17.8
seta_3	20	9.1	16.7	15.26666667
seta_4	14.3	0	14.3	9.533333333
seta_5	0	0	0	0
seta_6	25	12.5	12.5	16.66666667

Set B	TS1	(WER)		(Avg WER)
AudioClip		TS2	TS3	
setb_1	100	100	100	100
setb_9	100	100	100	100
setb_15	100	100	100	100
setb_19	100	100	100	100
setb_25	100	100	100	100
setb_30	100	100	100	100

Set C	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setc_2	16.7	50	41.7	36.13333333
setc_6	15.4	38.5	15.4	23.1
setc_7	0	0	0	0
setc_14	0	23.1	0	7.7
setc_18	12.5	62.5	0	25
setc_23	33.3	100	0	44.43333333

Set D	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setd_3	33.3	16.7	16.7	22.23333333
setd_5	0	28.6	0	9.533333333
setd_11	0	0	0	0
setd_14	0	0	0	0
setd_20	16.7	16.7	16.7	16.7
setd_27	20	0	0	6.666666667

Set E	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
sete_4	0	11.1	0	3.7
sete_8	33.3	50	16.7	33.33333333
sete_12	20	0	0	6.666666667
sete_17	0	0	0	0
sete_24	25	0	25	16.66666667
sete_29	16.7	33.3	16.7	22.23333333

Note: The audio clips are named as, set name followed by underscore and then person id.

Figure 1: The Group 1 Transcription Results

For Group 2

Set A	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
seta_7	0	0	25	8.333333333
seta_8	28.6	42.9	14.3	28.6
seta_9	10	4.1	10	8.033333333
seta_10	16.7	20	0	12.23333333
seta_11	33.3	16.7	16.7	22.23333333
seta_12	25	12.5	25	20.83333333

Set B	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setb_2	100	100	100	100
setb_7	100	100	100	100
setb_8	100	100	100	100
setb_16	100	100	100	100
setb_21	100	100	100	100
setb_27	100	100	100	100

Set C	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setc_1	16.7	10	30	18.9
setc_9	16.7	33.3	16.7	22.23333333
setc_12	20	20	40	26.66666667
setc_22	20	20	20	20
setc_28	20	20	20	20
setc_29	21.4	21.4	14.3	19.03333333

Set D	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setd_4	0	14.3	14.3	9.533333333
setd_6	18.2	27.3	18.2	21.23333333
setd_10	25	23.1	30.8	26.3
setd_13	25	25	25	25
setd_18	33.3	14.3	14.3	20.63333333
setd_30	20	40	20	26.66666667

Set E	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
sete_5	14.3	14.3	14.3	14.3
sete_11	25	25	25	25
sete_14	25	16.7	8.3	16.66666667
sete_15	16.7	33.3	16.7	22.23333333
sete_23	33.3	11.1	22.2	22.2
sete_26	14.3	14.3	14.3	14.3

Note: The audio clips are named as, set name followed by underscore and then person id.

Figure 2: The Group 2 Transcription Results

For Group 3

Set A	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
seta_13	30	20	9.1	19.7
seta_14	0	28.6	14.3	14.3
seta_15	14.3	14.3	14.3	14.3
seta_16	9.1	16.7	9.1	11.63333333
seta_17	11.1	11.1	9.1	10.43333333
seta_18	16.7	16.7	25	19.46666667

Set B	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setb_3	100	100	100	100
setb_10	100	100	100	100
setb_14	100	100	100	100
setb_20	100	100	100	100
setb_24	100	100	100	100
setb_29	100	100	100	100

Set C	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setc_5	14.3	14.3	0	9.533333333
setc_11	0	14.3	0	4.766666667
setc_13	42.9	42.9	14.3	33.36666667
setc_17	41.7	58.3	58.3	52.76666667
setc_21	60	60	30	50
setc_30	16.7	33.3	33.3	27.76666667

Set D	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setd_1	0	42.9	28.6	23.83333333
setd_7	7.7	46.2	7.7	20.53333333
setd_12	16.7	16.7	33.3	22.23333333
setd_15	0	16.7	16.7	11.13333333
setd_22	0	0	0	0
setd_26	0	33.3	66.7	33.33333333

Set E	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
sete_6	0	22.2	0	7.4
sete_9	0	0	0	0
sete_16	0	0	0	0
sete_19	14.3	0	14.3	9.533333333
sete_25	0	9.1	18.2	9.1
sete_28	50	33.3	16.7	33.33333333

Note: The audio clips are named as, set name followed by underscore and then person id.

Figure 3: The Group 3 Transcription Results

For Group 4

Set A	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
seta_19	6.7	0	0	2.233333333
seta_20	8.3	16.7	3.1	9.366666667
seta_21	22.2	0	0	7.4
seta_22	4.1	9.1	8.2	7.133333333
seta_23	27.3	36.4	10	24.56666667
seta_24	10.1	9.1	20	13.06666667

Set B	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setb_4	100	100	100	100
setb_6	100	100	100	100
setb_11	100	100	100	100
setb_13	100	100	100	100
setb_23	100	100	100	100
setb_28	100	100	100	100

Set C	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setc_3	0	28.6	15	14.53333333
setc_8	0	40	20	20
setc_15	12.5	25	20	19.16666667
setc_20	33.3	16.7	33.3	27.76666667
setc_24	9.1	9.1	20	12.73333333
setc_26	11.1	22.2	9.1	14.13333333

Set D	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setd_9	28.6	28.6	20	25.73333333
setd_16	16.7	16.7	33.3	22.23333333
setd_17	16.7	0	0	5.566666667
setd_19	0	0	10	3.333333333
setd_21	0	20	22.2	14.06666667
setd_25	22.2	22.2	9.1	17.83333333

Set E	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
sete_2	14.3	14.3	20	16.2
sete_7	0	9.1	10	6.366666667
sete_10	16.7	16.7	22.2	18.53333333
sete_18	0	7.1	9.1	5.4
sete_22	20	10	0	10
sete_27	0	0	20	6.666666667

Note: The audio clips are named as, set name followed by underscore and then person id.

Figure 4: The Group 4 Transcription Results

For Group 5

Set A	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
seta_25	0	27.3	18.2	15.16666667
seta_26	0	16.7	16.7	11.13333333
seta_27	21.4	28.6	14.3	21.43333333
seta_28	25	25	25	25
seta_29	0	0	12.5	4.166666667
seta_30	0	0	33.3	11.1

Set B	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setb_5	100	100	100	100
setb_12	100	100	100	100
setb_17	100	100	100	100
setb_18	100	100	100	100
setb_22	100	100	100	100
setb_26	100	100	100	100

Set C	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setc_4	0	0	0	0
setc_10	27.3	9.1	9.1	15.16666667
setc_16	50	50	50	50
setc_19	42.9	42.9	28.6	38.13333333
setc_25	0	40	0	13.33333333
setc_27	0	0	16.7	5.566666667

Set D	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
setd_2	0	20	20	13.33333333
setd_8	25	25	41.5	30.5
setd_23	14.3	14.3	14.3	14.3
setd_24	9.1	0	18.2	9.1
setd_28	0	0	20	6.666666667
setd_29	0	0	0	0

Set E	(WER)			(Avg WER)
AudioClip	TS1	TS2	TS3	
sete_1	23.1	30.8	30.8	28.23333333
sete_3	22.2	22.2	11.1	18.5
sete_13	14.3	14.3	20	16.2
sete_20	20	20	10	16.66666667
sete_21	28.6	0	14.3	14.3
sete_30	21.4	14.3	14.3	16.66666667

Note: The audio clips are named as, set name followed by underscore and then person id.

Figure 5: The Group 5 Transcription Results

VITA

Author's Name: Jayashri S. Chaudhari

Birthplace: Dhule, India

Birthdate: January 9, 1981

Education:

Bachelor of Science in Computer Science

North Maharashtra University, India

July 2002

Grade: First Class with Distinction (3.7/4.0)

Certifications

Sun Certified Java programmer 1.2, April 2004. Score: 96%

Professional/Research Experience:

University of Kentucky

Lexington KY

Aug 05- Aug 07

Graduate Research Assistant

D. N. Patel College of engineering, India

Department of Computer Engineering

Aug 02-Aug 03

Lecturer

Patents and Publications:

Chaudhari, J., S.-C. Cheung and M. V. Venkatesh. 2007. *"Privacy Protection for Life-Log Video"*, IEEE Signal Processing Society SAFE 2007: Workshop on Signal Processing Applications for Public Security and Forensics.

S.-C. Cheung and Jayashri Chaudhari, 2007. *"Audio-Visual Privacy Protection for Portable Video Recording Devices"*, United States Patent Pending.

Awards, Honors and Activities

Maharashtra State Talent Search Exam (MTS Exam), Consolation prize (1995).

Outstanding achievement recognition award, Maharashtra State (1996).

Reviewer for ICIP 2006, 2007, CVPR 2006, EURASIP.

Society Memberships

* IEEE Student Member