DoCoDeMo Phone: An Imperceptible Approach for Privacy Protection

Hsien-Ting Cheng, Ching-Lun Lin, Hao-hua Chu
Department of Computer Science and Graduate Institute of Networking & Multimedia
National Taiwan University, Taipei, Taiwan
{r92006, b89026, hchu}@csie.ntu.edu.tw

Abstract

Privacy protection on user context is one of the most important issues in ubiquitous and pervasive computing. This paper describes a new privacy protection approach that not only filters and reduces the granularity of context information to correct access level granularity, but also intelligently replace the filtered-out context information with the artificial context information considered appropriate by its user. The benefit of this new approach is that people who are accessing the filtered context information cannot detect if the context information has been filtered, i.e., filtering becomes imperceptible. Based on this new approach, we have designed, implemented, and conducted preliminary evaluation of the DoCoDeMo phone, which allows a user to imperceptibly conceal his/her surrounding ambient sound from the callers, without the callers knowing the presence of filtering.

1. Introduction

Do you sometimes feel that you are revealing more (context) information to others about your everyday life than what they really need to know? For an example, when we receive a voice call or participate in a video conferencing session, the information disclosed to our callers is not only our voice and facial expression, which are what we want them to hear/see. Since a microphone and a camera also capture and transmit ambient sound and background scene to our callers, these sensors can unintentionally reveal additional context information about our current locations or activities. In some situations, such additional information can cause unnecessary embarrassments and misunderstandings to the callees. To avoid these situations, we, as callees, often refuse to communicate to the callers at these inconvenient times. We provide the following two scenarios to illustrate these situations:

1. Joe told his girlfriend Jane that he was going to play basketball with his male friends. Unfortunately, some of his male friends did not show up for the game, so the basketball game was cancelled. Instead, Joe met some female friends and they decided to go to a coffee shop to have a friendly chat. At the coffee shop, Joe received a phone ring from Jane. Joe became hesitant to answer this phone call, because he was concerned that Jane could overhear the presence of his female friends, which might lead to unnecessary misunderstanding.
2. Joe noticed a video phone ring from his supervisor when he was entertaining an unexpected client at a local jazz bar. Since this was an unexpected visit, Joe did not inform his supervisor Jill about this. Again, Joe was hesitant to pick up the video phone call because he did not want Jill to oversee and overheard the bar and loud Jazz music and to think that Joe was slacking off from work.

A simple solution to this problem is to filter out the ambient sound and background scene [1]. However, this simple solution is *insufficient* in situations where the callers are *expecting* certain types of ambient sound or background scenes from the callees. In the 1st scenario, Jane would expect Joe to be in the basketball court and expect to hear the sound of playing basketball. In the 2nd scenario, Jill would expect Joe to be working in a busy office and expect to see/hear it. Ideally we want to *produce the expected ambient noise and background scene* – the basketball court or the busy office. Note that filtering alone can result in *noticeable absence* of ambient sound and background scenes, especially when callers are expecting callees to be in certain places with distinguishable ambient sounds and scenes. We believe that filtering may create an undesirable impression that callees are *intentionally and explicitly hiding* certain information from callers. Therefore, there is a need for a new approach in privacy protection that does not only protect callees' privacy on their context information, but at the same time, can make such filtering *imperceptible* to callers.

In this paper, we propose a new privacy protection approach that *not only* filter out context information, but also *intelligently substitute* the filtered-out context information with the *artificial context information* considered appropriate by its user, as to create the appearance of *imperceptible filtering* to the callers. Based on this new approach, we have designed an audio-based privacy protection system on a mobile device, called *DoCoDeMo*[1] *Phone*. To achieve imperceptible filtering, the DoCoDeMo phone is designed to do the followings: (1) it can filter out the background ambient sound from the callee's voice, (2) it can locate an appropriate ambient sound source expected by the caller, and (3) it can mix in a chosen ambient sound source with the callee's voice. In the 1st scenario described previously, the DoCoDeMo phone can filter out Joe's female friends' chatter in the background, locate an ambient sound source on basketball court, and mix it with Joe's voice. As a result, Jane can hear the expected location – the basketball court, where Joe is supposed to be. Joe can feel comfortable picking up phone calls anytime anywhere regardless of his current ambient environments.

2. Challenges

In order to realize the DoCoDeMo phone, we have identified the following technical challenges.

- *Detect if the user is at the location where he/she is expected.* This requires the user to maintain a schedule of what he/she is expected to be at different times. By looking up a user's schedule and comparing the expected location to his/her current location, the system can determine if the user is at the expected location. Currently, the most popular locating system is GPS. However, GPS does not work indoor. To solve this problem, the system pre-defines some location profiles, e.g., office, transportation, countryside, etc., and the user can manually change the current location profile. In addition, it is possible to apply audio recognition techniques on the ambient environment sound to automatically infer the current location profile of the user.
- *Quickly locate an appropriate ambient sound source at the expected location.* The amount of time to locate an appropriate ambient sound source must be no more than a few phone rings which is the amount of time the caller is willing to wait for the callee to pick up the call. If the appropriate ambient sound could not be identified in time, the caller might decide to disconnect the call and the callee would miss the phone call.
- *Filter out background ambient sound and mix the selected ambient sound in real time.* It is about how to efficiently reduce the background noise and mix new sound sources. Audio filtering and mixing have been active areas of research in speech processing. When choosing speech processing techniques for our system, we must consider the limited processing power on a mobile device as well as the real time constraints of a voice call.
- *Security attacks:* there are many possible attacks on our system to find out the location of a user. Consider the following attack. To find out whether a user (Joe) is at his expected location or not, an attacker can go to Joe's expected location and then make a phone call to Joe. If Joe is currently not at the expected location, Joe's cell phone will broadcast a request for the ambient sound source at the expected location. The attacker will then receive this request from Joe immediately after calling Joe. This means that the attacker can tell if Joe is at the expected place or not, depending on whether the attacker receives an ambient sound request at the expected place or not. Consider the 2nd attack. The attacker can monitor the data packet containing requests for ambient sound sources. Then the attacker can extract the IP address in the data packet and map it to the likely physical location.
- *Reliability:* an active ambient sound source can fail sometimes during a call. The failures can be caused by network disconnection from the unpredictable wireless networks, the source device running out of battery, etc. Since the failure of ambient sound source can make filtering visible to the callers, it has to be reliable under all these unexpected conditions.
- *Peer-to-peer architecture vs. centralized architecture:* there are two possible approaches to realize our system: peer-to-peer vs. centralized. We will compare the advantages and disadvantages of these two approaches.

3. Related Work

Previous works on protecting context information are focused mainly on information filtering and granularities. Project Aura [2] proposes an access control mechanism to filter out fine-grained information from rich, raw context data based on access privileges. For example, it is possible to determine location information from an image captured from a camera. If a user is granted only access to the location information, Aura will filter out and remove the image, and return only

[1] DoCoDeMo means everywhere in Japanese. That is, callee can appear to be anywhere to the caller.

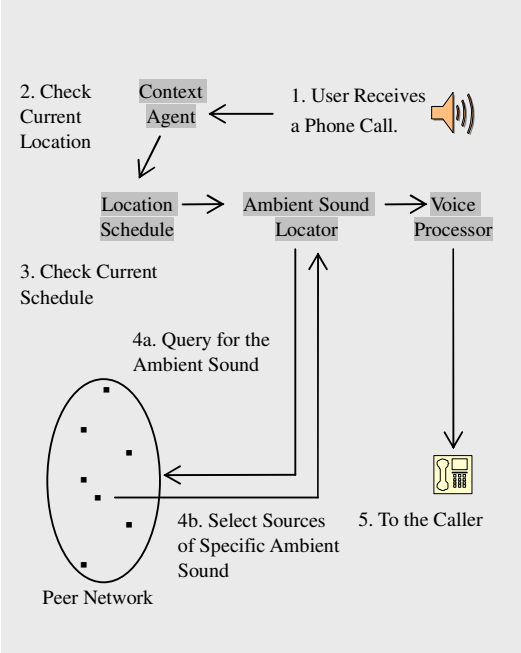

Fig 1. The executing flow of DoCoDeMo phone.

text-based location description. In comparison, our system does not only filter information granularity, but also intelligently substitute the filtered-out context information with the artificial context information. As a result, our system can create the appearance of *imperceptible filtering* to people who are accessing your context information.

There are several commercial products that can eliminate ambient noises. The Boom Noise Canceling Headset [1] allows users to communicate clearly in a loud noisy environment. This headset can be plugged into most cellular phones. It has two microphones. The *mouthpiece microphone* collects the user's voice with some of the ambient noises around the user. The *noise microphone* picks up all the ambient noise but little of the user's voice. The handset subtracts the ambient noise out of the noise microphone, from the audio signals out of the mouthpiece microphone. The net result of the subtraction is completely voice only.

Some cellular phone service providers, e.g., TransAsia telecom [3], currently offer services that allow a user to mix some background music into their phone calls. In order to use these services, a user must setup a schedule of when to mix the music or sound effect into the phone call. In addition, a user can decide what and how to mix the music or sound effect based on who the callee/caller is. Our system is different from these services on several aspects. Their motivation is based on making a phone call more *entertaining* rather than privacy protection. The selection of background music for mixing with voice is static and not context-aware as in our system.

4. Current Work

The design of our DoCoDeMo phone is shown in Fig. 1. It is consisted of four components: Context Agent, Location Schedule, Ambient Sound Locator, and Voice Processor. The executing flow is described using the following 5 steps:

1. Receive a phone call ring.
2. *Context Agent* gets the current location of the callee through GPS or the user's predefined location.
3. *Location Schedule* compares the callee's current location with his/her expected location schedule. If the callee is not at the expected location, prompt the callee to see if he/she needs ambient sound at the expected location.
4. *Ambient Sound Locator* finds several ambient sound sources and selects one as the active sound source.
5. *Voice Processor* filters out the original ambient sound and mixes in the ambient sound source.

We have implemented the DoCoDeMo phone on HP iPAQ running the Microsoft Windows CE Operating System. We have developed and deployed a voice processor that is capable of filtering and mixing ambient sounds. The ambient sound mixer is implemented by adding two waveforms and then adjusting the coefficients *a* and *b* to get the best performance.

$$S(t_k) = aS_1(t_k) + bS_2(t_k)$$

S_1 ,S_2: voice signal.; t_k: time index

Developing a smooth and effective ambient sound filter is more complex. It is equivalent to the problem of noise reduction in speech processing. We have found two general approaches for noise reduction. The 1st approach takes the whole voice sequence as an input, and then calculates a global optimal noise signal. This approach requires the knowledge of the whole voice sequence in advanced. The 2nd approach takes two frames at a time, and then calculates noise signals locally. This approach has an advantage of being applicable in frame by frame. Since our system needs to filter ambient sound frame-by-frame in real time, the 2nd approach is adapted. Among the frame-by-frame methods, we use the Wiener Filter, which is standardized by ETSI Technical Committee Speech Processing. [4]

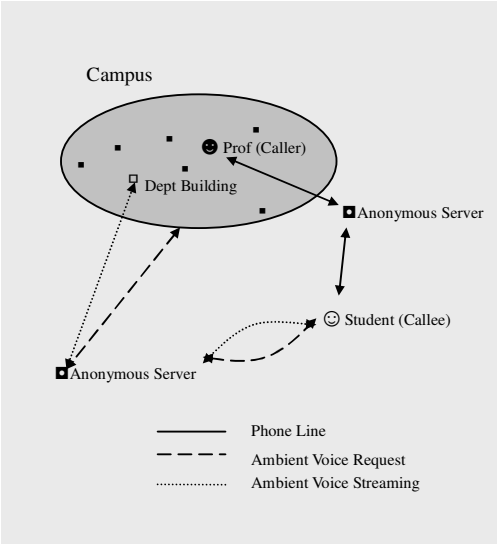

Fig 2. A professor wants to check if a student is currently working in the department building or is just fooling around. If the student is not at his position, the student can ask for an ambient sound source provided by other students on campus through anonymous server.

Noise type	Speech		Jazz		Rock music	
Input SNR [dB]	5	-5	5	-5	5	-5
Improve	7.3	11.6	8.9	13.3	8.2	12.7

Tab 1. The improvement of the noise reduction using Weiner Filter for three different noise type: speech, jazz, and rock music. And each type with two input SNR (in dB).

Testing Item	Performance
Mix Quality (score)	1.....2.....3.....4.....5

Tab 2. The average mix quality: the score is made from 1 to 5, the bigger the better. We got 4.2 in our subjective evaluation.

	None	filter only	Both filter &mix
Delay time (sec)	0.64	1.53	1.71

Tab 3. Delay time (process + transmit): evaluated with none of filter and mixer, filter only, and both filter and mix respectively.

5. Proposed Design

There are two possible approaches to realize our system: centralized vs. peer-to-peer (P2P). In a centralized architecture, the system maintains a directory server of clients' locations, which periodically pulls clients' locations (or it can be client-push to the location matching server). The location directory server responds to clients' requests for the ambient sound source at a specified location. The centralized architecture can also deploy powerful, stationary *voice processing servers* to run audio filtering and mixing software, so they can alleviate the problem of limited processing and power on a mobile device. In the centralized architecture, all voice and data communications must go through some centralized servers that are responsible for the privacy and security protection of users making requests or providing ambient sound sources.

However, the centralized architecture requires infrastructure support to deploy this service. Therefore, we favor the P2P architecture, i.e., direct peer node communications between the requestors and providers of ambient sound sources [5]. We prefer the P2P architecture because our application is fundamentally of *a P2P flavor* – where the participating users are acting as *service providers* sharing their ambient sounds and also as *service consumers* when they are in need of ambient sounds.

We propose using *anonymous redirection server*s to address the security attacks described in section 2. Recall that the security attack is about the attacker who can monitor the data packet containing requests for ambient sound sources, and then infer physical location from the IP address in the data packet. Using anonymous servers stops this attack by having the requests redirecting through anonymous servers; therefore the anonymous servers can remove the IP address from the request. This is shown in Fig 2.

To increase the reliability of ambient sound sources, we apply redundancy in ambient sound sources. That is, the system will select one active ambient sound source and maintains several sound sources as backup in case the active source fails.

6 Preliminary Evaluation

We have conducted preliminary evaluation on the performance of DoCoDeMo phone. The evaluation metrics include the quality of ambient sound filtering, the quality of ambient sound mixing, and communication delay. In general, speech quality evaluation can be done objectively or subjectively. We apply objective evaluation in ambient sound filtering and subjective evaluation in ambient sound mixing.

- *Ambient sound filtering:* To measure the noise reduction performance of Weiner Filter, we use segmental SNR (signal-to-noise ratio) improvement. It is calculated by:

$$SNR_{improve} = segSNR_{out} - segSNR_{in}$$

  Three types of background noises (speech, jazz, rock music) are artificially added to a clean speech. Each noise has different SNR: -5dB and 5dB. The results of noise reduction are shown in Tab 1.
- *Ambient sound mixing:* we asked a total of 25 peoples to score the quality of our ambient sound mixer. The score can have values ranged from 1 to 5 with the higher value the better. The results are shown in Tab 2 with the average score of 4.2.
- *Communication delay:* we measure the communication delay time, which includes the processing times for audio filtering and mixing, and network delay. Our preliminary results in Tab 3 have shown good performance in voice processing time and network delay time. Experiments have been conducted under three different conditions: without filtering and mixing, with filtering only, and with both filtering and mixing.

7 Future Work

As shown in Tab 1, the quality of ambient sound filtering in speech is not as good as in noise. We believe that it may be due to the fact that the difference between user's voice and background speech voice is not significant enough to distinguish them. People's conversation on the background sometimes shows more information than other ambient sound. We would like to enhance the filter so that it can fully detect and remove the background speech.

Our current system is designed and implemented based on the P2P architecture. We would like to improve this P2P application in its scalability, security and privacy, quality of services, performance, fault tolerance, etc.

We believe that this new privacy protection approach can be easily applied to video that can have sensitive background scenes that the user may not want to share with the callers. We are looking for video filtering and mixing methods that can change the background scenes without being noticeable to the callers.

References

[1] "The Boom Noise Canceling Headset" http://www.thetravelinsider.info/roadwarriorcontent/boomheadset.htm

[2] *U. Hengartner, P. Steenkiste,* "Access Control to Information in Pervasive Computing Environments", HotOS, 2003.

[3] TransAsis Telecommunicatios http://www.hank.net.tw/channel/TL/BGM/service.htm

[4] *ETSI ES 202 050 v.1.1.3 (2003-11),* "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms". 5.1 Noise Reduction.

[5] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker. "Search and replication in unstructured peer-to-peer networks,". *Proc. of the 16th ACM Int'l Conf. on Supercomputing, June 2002.*