Sound Shredding: Privacy Preserved Audio Sensing

Sumeet Kumar , Kate Liu , Pang Wu and Ying Zhang

Carnegie Mellon University, Moffett Field, CA 94035

{sumeet.kumar, kate.liu, pang.wu, joy.zhang} @sv.cmu.edu

ABSTRACT

Sound provides valuable information about a mobile user's activity and environment. With the increasing large market penetration of smart phones, recording sound from mobile phones' microphones and processing the sound information either on mobile devices or in the cloud opens a window to a large variety of mobile applications that are context-aware and behavior-aware. On the other hand, sound sensing has the potential risk of compromising users' privacy. Security attacks by malicious software running on smart phones can obtain in-band and out-of-band sound information to infer the content of users' conversation. In this paper, we propose two simple yet highly effective methods called *sound shredding* and *sound subsampling.* Sound shredding mutates the raw sound frames randomly just like paper shredding and sound subsampling randomly drops sound frames without storing them. The resulting mutated sound recording makes it difficult to recover the content of the original sound recording yet we show that the acoustic features are preserved which do not affect the accuracy of activity and context recognition. Through user study, we demonstrate that this simple technique of sound shredding preserves privacy without sacrificing the efficacy.

Categories and Subject Descriptors

K.6 [**MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS**]: K.6.5Security and Protection

General Terms

Security and privacy of mobile computing

Keywords

Mobile sound sensing, sound sensing, sound shredding, sound subsampling, sound randomization, user privacy, human activity recognition and context recognition.

1. INTRODUCTION

Mobile sound sensing, which uses acoustic attributes collected by mobile devices has been found useful in diverse scenarios. Because audio data may contain unique fingerprints, allowing sound sensing software to extract and recognize meaningful events, many applications and systems have already applied sound sensing to improve their approaches . For instance, SurroundSense [2] uses acoustic and other attributes to identify user motions and SensOrchestra [6] leverage sounds and images to recognize the location form where those data were collected. These research results demonstrate that sound sensing could be of significant value in context and activity recognition as well.

As illustrated in Figure 1, in a typical audio based application, sounds are collected by mobile devices (either phones or tablets), and stored in storage like SD cards. These mobile devices are usually equipped with high sample rate microphones, which is useful for audio-based applications such as audio recording, speech recognition, and sound sensing. However, the benefit entails the risk of privacy when it comes to collecting audio data. The raw audio data from the microphone are insecure and could easily be replayed. Also sensitive informations are occasionally communicated over phone because audio is generally considered more secure than other mediums like text and email. However, emails and SMS texts are often encrypted by the applications while storing them in the mobile phone, which is often not the case with audio data collected by sound sensing applications. Because many of the sound sensing application collects continuous audio data which are rarely encrypted, the audio data could record personal conversations as well as sensitive informations. The replayed sound, even at a low sampling rate, may reveal the identity and other sensitive information of the users. Thus the raw sounds may be abused to disrupt the privacy guarantees for users. The problem becomes more obvious in case of continuous sampling applications like MobiSens [10] which uses the audio data for human activity recognition.

To mitigate this privacy concern, this paper proposes two approaches, sound shredding and sound subsampling, which are both aimed at preserving user privacy without overly compromising the accuracy of sound sensing. To achieve this goal, we preprocess the audio recordings, either using sound shredding or sound subsampling, before storing them.

The idea of sound shredding is to prevent the replay of audio recordings. To do so, we shuffle the snippets of an audio recording in a random order. By reordering the snippets, this approach can hinder malicious software from com-

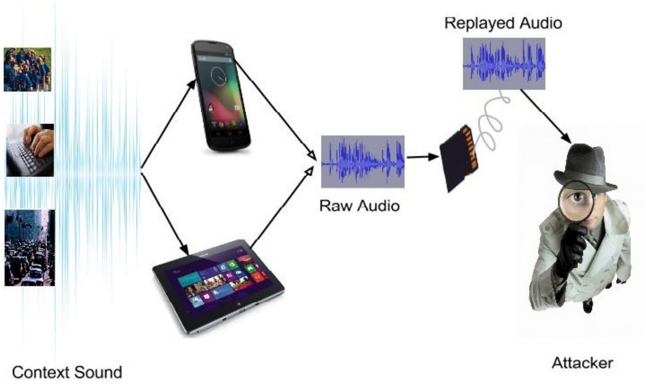

Figure 1: Insecure sound collecting process: sounds are collected by phones or tables, and stored in external storage like SD cards. The sounds are not encrypted and may be stolen by malicious software.

promising user privacy. Since the recordings are distorted, the attacker won't be able to understand the conversation or identify the participants. In addition, we provide another method, sound subsampling, which drops parts of the snippets from the audio recording. This method can also achieves the similar goal as sound shredding does.

While the two approaches delimit the attackerâĂŹs abilities, they still provide sufficient fingerprints for sound sensing software to extract useful events. The results in section 4 shows that sound shredding and sound subsampling can mitigate the threat from malicious software that intends to sniff information from the raw sounds. Also, the accuracy of context recognition does not significantly drop using Subsampling and the results of activity recognition using our approach, stays intact with sound shredding.

The main contributions of this paper are:

- **Methods to preserve user privacy:** We address the concern of privacy guarantees that may be undermined by malicious software intending to sniff information from raw sounds. User privacy can be preserved by preprocessing raw sounds with sound shredding and subsampling. Meanwhile, the recognition accuracy would not be overly sacrificed.

- **Experiments and evaluation of sound shredding and sound subsampling:** The goal of the two proposed methods is to preserve the user's privacy without significantly decreasing accuracy. Therefore, we compare the results using raw sounds to the ones using sound shredding and sound subsampling in section 4. We also quantify our findings in this section.

The rest of the paper is organized as follows. We discuss related work in section 2. Our sound shredding and recognition methods are described in section 3. In section 4, the experiments and results are elaborated and evaluated. We conclude our work in section 5.

2. RELATED WORK

Sound sensing is commonly used in context recognition. In the field of location recognition, SensOrchestra [6] achieved 87.7% recognition accuracy using audio and image attributes collected by mobile phones. This method groups nearby phones and integrates the estimates from sounds and images to recognize a location. In addition to human activity recognition, sound can also be used to determine emotion changes. For instance, StressSense [8] used human voice recorded by smartphones to recognize stress. The authors demonstrated that their method could robustly identify stress across multiple individuals in diverse acoustic environments. The method achieved 81% and 76% accuracy for indoor and outdoor environments, respectively. Those experimental results show the usefulness of acoustic attributes.

In addition, speaker recognition technology has been developed and discussed for long. Various applications have already applied speaker recognition to provide contextual information. Specifically, in the work of SoundSense [9], the authors were able to discover sound events to individual users. Their sound learning algorithm learns a unique set of acoustic events for each individual user and therefore provide personalized context. Although the algorithm improves the usability of mobile applications that tend to provide custom information, it also brings privacy implications.

Unfortunately, those sound sensing methodologies discussed above did not take the privacy implications into account, therefore introducing potential attacks against user privacy. Although there is a plenty of research on using audio sensing for diverse applications, not much has been done on securing the collected audio data on mobile phone. There are many encryption techniques available to secure any data but the limitations on mobile phones demands a technique which could easily be implemented and does not consume much power, even if the application runs continuously. Also, the technique should preserve the acoustic features that are used by the sound sensing applications. Through our experiments, we show that some light-weighted techniques are possible which can be executed on a mobile phone and still be effective in improving the privacy of the users.

3. METHODOLOGY

A variety of mobile applications collect data from sensors such as accelerometers, GPS, audio etc and use the sensor data for human activity recognition. One of such applications is MobiSens [10]. MobiSens uses adaptive activity recognition that not only builds model though automatic segmentation but also adapts the model over time. MobiSens had earlier used accelerometers, GPS and light as modalities for activity recognition, but the recent experiments have shown sound sensing to be an important modality in the supervised activity recognition.

Since mobile sensing applications like MobiSens collect continuous data from users, privacy is a genuine concern. The concern is more obvious in case of audio data because audio data could easily be replayed. A user is often comfortable with continuous collection of the accelerometer data but not with the audio data. Although sophisticated techniques to secure data are available, applying them on a mobile device consumes power. In this section, we elaborate the process of context and activity recognition using audio data. In addition, we also propose two ways to improve the users privacy while collecting audio data, namely ”Sound Sub-sampling” and ”Sound Shredding” which are novel approaches to collect audio data in a format that can improve user privacy without consuming much of the phones battery as consumed in many encryption and decryption technologies. Sound Sub-Sampling improves the privacy by storing the minimum raw data from which the context information could be extracted later, whereas, Sound Shredding randomizes the audio frames but still keeps the activity recognition percentage intact.

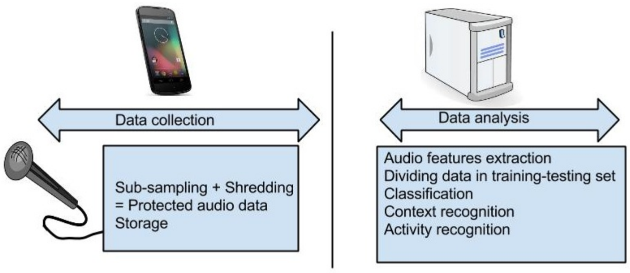

Figure 2: System Architecture: Sound is recorded using a mobile phone and in the process of recording, the audio data is shredded and sub-sampled for improved privacy. The data is later sent to server for analysis.

Figure 2 shows the architecture of the system. Sound is recorded using mobile phone and is secured by sub-samping and shredding. The secured audio data could be later sent to server for analysis.

3.1 Context and Activity Recognition Using Audio data

Audio data have been used for both context [7] and people centric activity recognition [9]. We define context as background in which an activity happens. For example, when a person is taking out money from an ATM, taking out money is an activity whereas ATM room is the context. Similarly, when a person is having dinner in a restaurant, dining is an activity whereas restaurant is the context. The process of context and activity recognition involves these steps:

- **Audio data collection:** Audio data could be collected using any device with a microphone. For our experiment, a total of thirty two sounds samples were recorded with a sampling rate of 8KHz sampled at 16 bit using a Nexus 4 phone. Each of the recordings were in the range of 2-3 minutes. We paid attention to collect data from a variety of environments encountered in day to day life.

- **Features extraction:** First, the audio data is framed using a sliding window with window size of 30 ms and with a 50% overlap. Then hamming filter is used to smoothen the window. After that, for each of these audio frames, Mel-frequency cepstral coefficients (MFCC) of 12 vector length are calculated. Mel-requency cepstrum (MFC) is a representation of power spectrum of sound which are widely used for many sound applications that need to compare audio in human audible range. It is calculated by linear cosine transformation on power spectrum on mel scale of frequency which approximates auditory system response and hence is better suited for comparing audible sound than other scales.

- **Context Recognition:** The experiment uses two machine learning algorithms for context classification. In this paper, we present the results of K Nearest Neighbor (KNN) and Support Vector Machine (SVM).

  KNN is a simple but yet very versatile algorithm used in many data mining applications. We used Java-ML [1], a widely used library machine learning and data mining algorithms, which provides a simple interface and an easy way to compare different classifiers. In KNN, the value of k determines how many neighbors influences classification. For our experiment we used k=5 to find five closest MFCC vectors.

  The other algorithm used is SVM, which is another popular learning algorithm often used for supervised learning models as demonstrated by the work of [4]. SVM is a discriminative classifier which is based on finding a hyper plane that gives largest minimum distance to the input points. We used LIBSVM which is a library of support vector machines [3] algorithms popularly used for solving classification and optimization problems.

  The MFCC features extracted from the audio data were randomly divided in two groups in the ratio of 8:2 for training and testing purpose. The training and testing data were used as input to the above two algorithms for the context recognition accuracy. We again used Java-ML for running experiments, which provides an easy interface to get classification results which are presented in the next section.

- **Activity Recognition**. In most cases, sound information associated with activities is not homogeneous as “coffee grinding” and “mowing lawn”. Rather, to describe the activity we need to model the acoustic characteristics of the event over a much longer time span. For example, activity “withdrawing cash at ATM” involves sound of “push buttons on keypad”, “ding sound from ATM”, “cash dispenser” and other clips. In other words, to recognize activities we need to go beyond acoustic frame level and model longer span to recognize the high-level meaning of the acoustic signal timeseries.

  In this paper, we adopt a simple method of modeling sound sequence as acoustic “sentences” as illustrated by Chen [5] for accelerometer readings. The main idea is to convert each sound sample to an acoustic “sentence” and use the acoustic word frequency distribution to characterize different activities. Note that the *frequency* of acoustic words distribution is the term used in statistical language processing describing how often a word type occurs in a sentence or a document rather than the physical frequency of sound vibration.

  We quantize the MFCC features in V groups using K-Means clustering algorithm. Once the K-Means clustering algorithm converges, we get V cluseters. We give each of the cluster a unique label and use he same label to mark the MFCC feature vectors based on the cluster it falls in. Then, we use a sequence of labels as a representation of an activity. Similar activities tend to have similar representation, which also means that they have same frequency of label distribution and hence similarity between activities could be predicted by comparing frequency of labels in different activities.

3.2 Sound Subsampling

Identifying speech from an audio source requires a fairly continuous data but that is not the case with context recognition. The context of an activity can often be extracted from a few segments of sound e.g. if a person is driving his car as well as talking to his fellow rider, the extraction of speech requires a continuous sample whereas the background noise of a moving car on the road can be extracted from even a few audio segments. In fact, the context recognition like driving a car does not require a continuous sample. At the same time if continuous sample is not collected, it makes it difficult to retrieve speech information. Hence, if context recognition is the primary goal, users privacy could improve by storing sub-samples of audio data rather continuous samples of audio data.

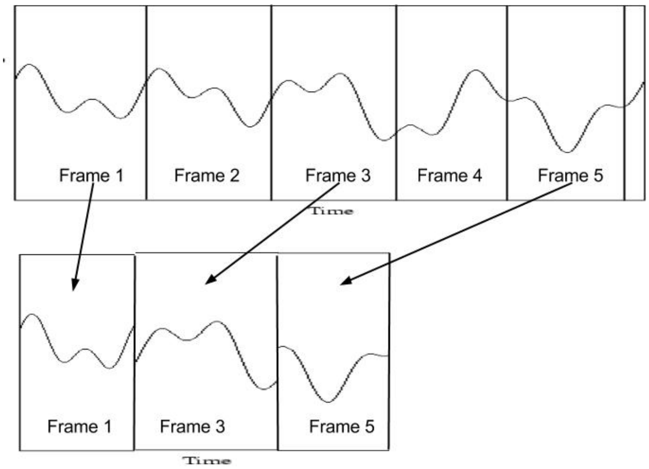

Figure 3: Sound Sub-sampling at the rate of 50%

We define Sub-sampling as the process of collecting a part of the raw data e.g. a subsampling of 50% means only 50% of audio data is stored, i.e. one audio frame is discarded after every single audio frame stored. Figure 3 demonstrates the process of sub-sampling where every second frames is being dropped during the audio data collection process.

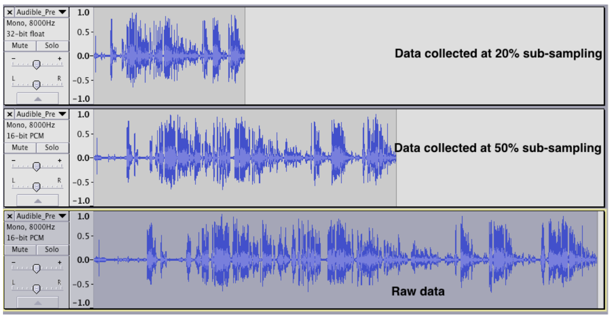

Figure 4: Sound Sub-sampling at the rate of 20%, 50% and Raw Data

We collected audio data for our experiment with around thirty sound samples having different backgrounds like driving, walking, meeting etc. For each of these sound samples we designed experiments to get sub-sampled data at different rates of sub-sampling, to find an optimized ratio of subsampling for privacy preserved audio data collection. Figure 4 shows that the original sound and the sub-sampled sound have a similar looking shape because even the sub-sampled sound contains the original sound signature.

Later we used the subsampled data to find their context recognition accuracy. We also conducted a user study to find any effect on the privacy by sub-sampling the raw audio. Our study revealed there is a drop in context recognition with sub-sampling but, even after dropping around 70%-80% of audio data, there is not a major drop in context recognition accuracy. Hence an application which targets context recognition can tread some recognition accuracy for a much improved user privacy.

3.3 Sound Shredding

The subsampling of audio is good way to reduce speech information in the audio data, but even sub-sampling at a lower rate could still give away users information like number of people in conversation or the gender for the person speaking etc. One possible way to further improve user privacy is by randomizing the sound data. We noticed that sound features like MFCC are extracted from audio frames of 30 ms duration. These features do not change even if the sound frames are randomized as long as the frames are not changed internally. Figure 5 shows the process of frames randomization where the frames are randomly moved to a another location in the audio frames sequence. At the same time, if sound frames are randomized it becomes difficult to reconstruct and replay the sound. We used this for improved privacy during audio data collection.

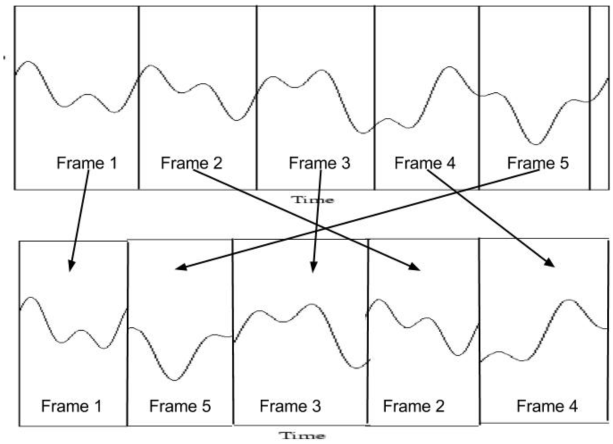

Figure 5: Sound shredding

We define Sound Shredding as randomizing the audio frames in a sound snippet. We randomize sound by selecting an audio frame and moving it to a random location in the sound snippet i.e. if a frame is located at i index in the collection of audio frames that makes the sound snippet, we generate a random number between 0 and i, and move the frame at the generated random number. We do the same with all the frames that make the sound snippet.

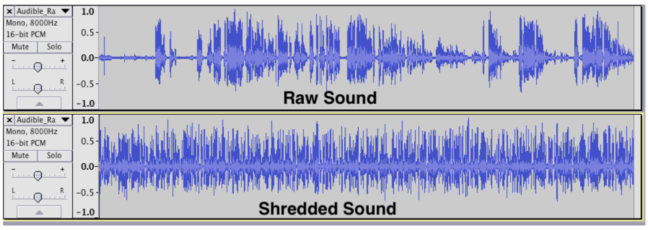

Figure 6: Sound Shredding: Raw data and shredded data

Figure 6 shows the data collected by shredding. As the data is randomized during collection, the shredded data looks very different from sub-sampled data.

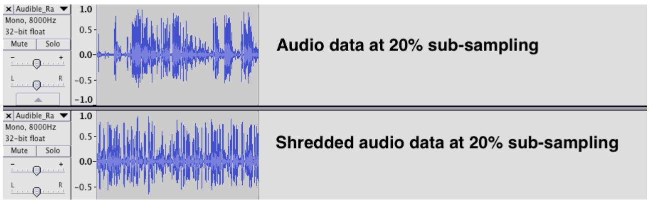

Figure 7: Sound Shredding and sub-sampling: Sub-sampled (20%) data and Sub-sampled (20%) shredded data

3.4 Sound Shredding and Sound Sub-sampling

Though Sound Shredding and Sound Sub-sampling have their own usage, the two can be combined in some cases for improved data privacy. To combine sound shredding and sub-sampling, the audio data being collected is first randomized (shredded) and then sub-sampled which results in shredded sub-sampled data.

Figure 7 shows the data collected by sub-sampling at 20% and the data collected by shredding and sub-sampling at the rate of 20%.

4. EXPERIMENTS AND RESULTS

In this section, we describe our experiments that are divided in two parts. We conducted experiments to determine the effect of sound sub-sampling and sound shredding on context and activity recognition accuracy. We also conducted a user study to find the changes in user privacy with sound shredding and sub-sampling.

4.1 Experiment Setup

Audio data for the experiments was collected using a Nexus 4 phone by reading its microphone at 8000 HZ using single audio channel. The experimenter collected thirty-two sound samples in different contexts as described in Table 1.

The experimenter used the context as the label for the audio. For each of the contexts, four sound snippets of approximately 2 minutes duration were recorded.

4.2 Context Recognition

We divided the raw audio snippet in frames of 30 ms, which were used to extract the MFCC(12) features. For testing the algorithms accuracy, we divide the entire set of MFCC features in to training and test data. We used 80% of the data set as training data and the rest as test data. To classify the context, we used proven KNN and SVN algorithms. A collection of vectors made of 12 coefficients of MFCC and the audio label was used as input to the classification algorithms. The training and testing data were used as input to the above two algorithms for the context recognition accuracy. We used Java-ML [1] for running experiments, which provides an easy interface to get the classification results.

Table 1: Sound Contexts

No	Context	No of Recordings
1	Student Faculty meeting	4
2	Friends talking during lunch	4
3	Brewing coffee in cafeteria	4
4	Students talking in a meeting	4
5	Walking on road	4
6	Classroom discussion	4
7	Guest talk in a conference room	4
8	Doing experiments an a laboratory	4

The experiments were run with varying degree of sub-sampling.

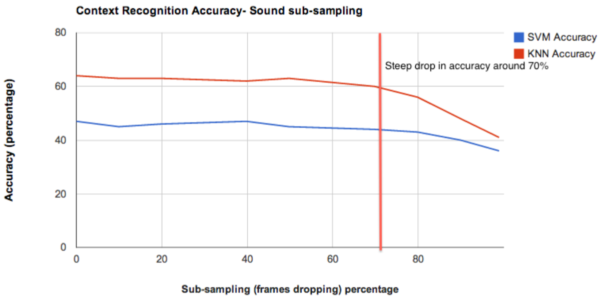

Figure 8: Context Recognition Accuracy vs. Sound sub-sampling percentage

Figure 8 shows the trend of change in accuracy of SVN and KNN algorithms with change in sub-sampling percentage. The results show a slow decrease in recognition accuracy with increased sub-sampling (increased frames dropping) till the sub-sampling percentage is around 70%. But after 80% sub-sampling there is a siginificant decrease in context recognition accuracy.

In addition, the experiments were also run with sub-sampled shredded sound with varying degree of sub-sampling.

Figure 9 shows the trend of change in accuracy of of SVN and KNN algorithms with change in sub-sampling percentage for shredded audio. The results show a slow decrease in recognition accuracy with increased sub-sampling (increased frames dropping) till the sub-sampling percentage of 80%. But after around 80% sub-sampling, there is a siginificant decrease in the context recognition accuracy.

The above experiments and results show that shredding and sub-sampling of audio data can lead to improved data privacy without losing much on recognition accuracy, if sub-sampling and shredding has positive impact on users privacy. The impact of sub-sampling and shredding on users privacy

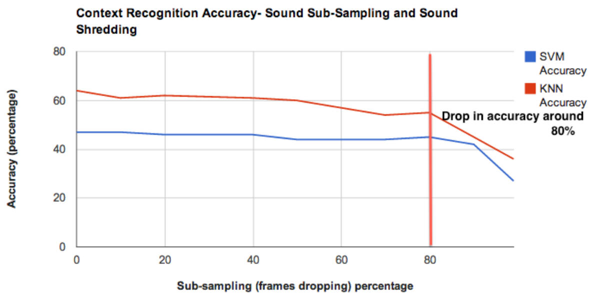

Figure 9: Context Recognition Accuracy vs. Sound sub-sampling percentage for shredded sound

is being discussed in the coming sections.

4.3 Activity Recognition

In this experiment, We use activity as a language approach of activity recognition as explained in [?]Cheng:2010:LAI:1891903.1891937). After data collection and feature extraction, we cluster the MFCC vectors using k-means clustering, each cluster with their own labels. This results in a representation of the sound snippet as sequence of labels. The sequence could consist of a number of activities. Later we randomly extract a continuous segment of around 20% labels as testing data, assuming it to be one activity and the remaining data is used for training algorithm. For each of the test data cosine similarity is used based on frequency of labels to find top few similar activities.

In the training phase:

- Convert sound frame to acoustic words. The first step is to convert each sound frames from their MFCC vector representation to acoustic "words". We run K-means clustering over all observed sound frames and cluster these MFCC vectors to K clusters. Each sound frame is then replaced by the its corresponding cluster labels. For sound frames in the testing data, its MFCC vector is compared with each of the K cluster centroids and then labeled as the cluster nearest to.
- Convert a sound segment to an acoustic sentence. For example, the sound segment of "withdrawing cash from ATM" can be represented as a string of "N N T T T N T N T T T T T T N T T N T N T T T T T N N N N N N N P N N N N T T N."
- Modeling activities by their acoustic word frequencies. We use the normalized word frequency to represent each activity. For example, "withdrawing cash from ATM" has the acoustic word frequency vector as (T 0.50, N 0.47, P 0.03) which reflects the distribution of different acoustic words in this activity.

In the testing phase:

- Convert sound frame to acoustic words. For each sound frame's MFCC feature vector, calculate its distance each of the K cluster centroids. Label the sound frame using the cluster ID with the shortest distance.
- Convert the testing sound segments to acoustic word frequency vector such as (T 0.32, N 0.23, P 0.15, GB 0.1, FF 0.1, X 0.1).

VocSize(K)	Top 1	Top 2	Top 3	Top 4	Top 5
10	0.26	0.45	0.60	0.72	0.78
20	0.32	0.54	0.66	0.75	0.82
30	0.28	0.51	0.65	0.73	0.79
40	0.33	0.56	0.70	0.80	0.84
50	0.42	0.65	0.76	0.84	0.89
60	0.31	0.52	0.66	0.75	0.81
100	0.30	0.51	0.70	0.81	0.86
200	0.28	0.44	0.59	0.72	0.81

Table 2: Activity recognition accuracy using raw audio. Each row shows the setting of different sound word vocabulary size which corresponds to K in K-means clustering of sound frame MFCC vectors. "Top 1" is the percentage of true activity being predicted as the top-1 activity by the sound recognition algorithm. Similarly, "Top 2" shows the percentage of testing sound segments having their true activity labels being predicted in the top 2 most likely activities.

- Calculate the distance of the testing segment's acoustic word frequency vector with that of each activities in the training data and label the testing data as the activity that is closest.

The sound shredding method proposed in this work only randomly shuffles the sound frames which only affects the order of different frames and does not change the distribution of acoustic word frequencies. Because sound shredding uses acoustic word frequencies to compare activities, which does not change with sound shredding, Sound Shredding does not impact the activity recognition accuracy. On the other hand, the subsampling technique removes frames from the raw recording and has the potential risk of altering the distribution of acoustic word frequencies.

Table 4.3 and 3 list activity recognition accuracy using raw (or shredded sound) vs. using 20% subsampled sound data. Notice that, in these two results, we only use sound information for activity recognition and the training data is much smaller compared to results reported in the literature. The key point here is that by comparing the results using raw audio vs. using subsampled sound, there is a significant drop in activity recognition. Thus, if the sound sensing applications need to maintain high activity recognition rates, sound shredding is preferred than sound subsampling.

4.4 Privacy Preservation User Study

To find any impact on data privacy, we shared the raw audio data as well as audio data with sub-sampling and sound shredding. Then, we conducted a user study with students. The user study involved playing different sounds in front of users. As they hear the sound, they rated the sound on speech recognition, finding the number of people in conversation and gender identification. The scale used was 1-5, where 1 meant ”Not at all” and 5 meant ”Yes, I can”. Over all, 10 people took the survey and the responses were averaged to use in the graph. Parameters and scale used for user study:

1. Speech recognition (1- 5)
2. Count of people in conversation (1-5)

VocSize(K)	Top 1	Top 2	Top 3	Top 4	Top 5
10	0.19	0.34	0.48	0.59	0.69
20	0.22	0.43	0.57	0.68	0.73
30	0.16	0.34	0.49	0.58	0.66
40	0.19	0.41	0.55	0.64	0.71
50	0.28	0.51	0.63	0.73	0.80
60	0.16	0.35	0.50	0.59	0.65
100	0.18	0.29	0.41	0.48	0.59
200	0.10	0.23	0.33	0.41	0.50

Table 3: Activity recognition accuracy using 20% subsampled sound (dropping 80% frames). Each row shows the setting of different sound word vocabulary size which corresponds to K in K-means clustering of sound frame MFCC vectors. "Top 1" is the percentage of true activity being predicted as the top-1 activity by the sound recognition algorithm. Similarly, "Top 2" shows the percentage of testing sound segments having their true activity labels being predicted in the top 2 most likely activity

3. Gender identification (1- 5)

The data obtained was aggregated in a chart format shown in Figure 10. As it can be observed the speech recognition, one of the major concern is user privacy significantly improves by sound shredding in which audio frames are randomized. In addition, the possibility of counting people decreases with shredding as well as sub-sampling. The gender identification showed least improvement, but still improves by 10-25%. Overall sound shredding with subsampling rate of 20% gives the best result in terms of privacy preservation.

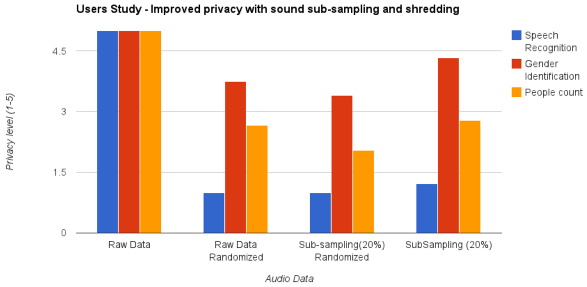

Figure 10: The user study results indicate that sound shredding can effectively protect user privacy. The speech recognition which is the primary means to obtain sensitive informationm, decreases significantly by using our approaches.

5. CONCLUSION

The experiments on sound shredding and sound sub-sampling when combined with user study results, gives a clear indication that user privacy significantly improves with sound shredding and sub-sampling. Because sound shredding does not change the audio frames internally, there was no significant difference in context and activity recognition with sound shredding, but the user privacy significantly improves. Also sound sub-sampling (till 70%) has minor decrease in context recognition accuracy but great improvement in user privacy. Both sound shredding and sound sub-sampling are effective ways to improve audio data privacy, but their usage would depend on the need of the application.

Although, sound sub-sampling and sound shredding can have significant impact users privacy when used in audio data collection, they are not full-proof means to protect users privacy. Given enough computation power there is a possibility of reproducing some part of the original audio.

6. REFERENCES

[1] T. Abeel, Y. Van de Peer, and Y. Saeys. Java-ml: A machine learning library. *J. Mach. Learn. Res.*, 10:931–934, June 2009.

[2] M. Azizyan, I. Constandache, and R. Roy Choudhury. Surroundsense: mobile phone localization via ambience fingerprinting. In *Proceedings of the 15th annual international conference on Mobile computing and networking*, MobiCom '09, pages 261–272, New York, NY, USA, 2009. ACM.

[3] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[4] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon. Large-scale content-based audio retrieval from text queries. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 105–112, New York, NY, USA, 2008. ACM.

[5] P.-W. Chen, S. K. Chennuru, and Y. Zhang. A language approach to modeling human behaviors. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010.

[6] H.-T. Cheng, F.-T. Sun, S. Buthpitiya, and M. Griss. Sensorchestra: Collaborative sensing for symbolic location recognition. In M. Gris and G. Yang, editors, *Mobile Computing, Applications, and Services*, pages 195–210. Springer Berlin Heidelberg, 2012.

[7] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):321–329, 2006.

[8] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury. Stresssense: detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 351–360, New York, NY, USA, 2012. ACM.

[9] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, MobiSys '09, pages 165–178, New York, NY, USA, 2009. ACM.

[10] P. Wu, J. Zhu, and J. Y. Zhang. Mobisens: A versatile mobile sensing platform for real-world applications. 18, February 2013.