Sound Shredding: Privacy Preserved Audio Sensing

Sumeet Kumar, Le T. Nguyen, Ming Zeng, Kate Liu, Joy Zhang
Carnegie Mellon University
Moffett Field, California, USA
{sumeet.kumar, le.nguyen, ming.zeng, kate.liu, 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 text content of the original sound recording, yet we show that some acoustic features are preserved which retains the accuracy of context recognition.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.5.5 [**Information interfaces and presentation (e.g., HCI)**]: Sound and Music Computing

Keywords

Sound sensing; sound shredding; sound subsampling; user privacy; context recognition

1. INTRODUCTION

Mobile sound sensing, which uses acoustic attributes collected by mobile devices has been found useful in diverse scenarios of context awareness. 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 [4] leverage sounds and images to recognize the location from where those data were collected. These research results clearly demonstrate that sound sensing could be of significant value in context recognition.

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 are useful for audio-based applications such as phone conversation, speech recognition, and sound sensing etc. 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. The replayed sound, even at a low sampling rate, may reveal the identity and other sensitive information about 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 such as MobiSens [13].

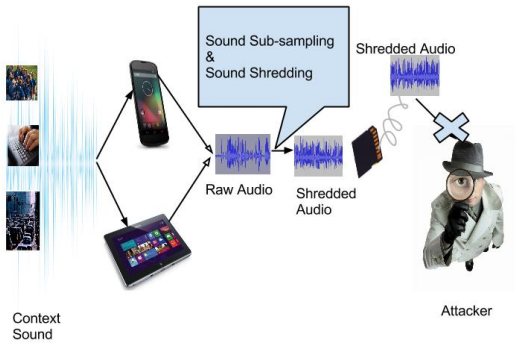

Figure 1: Shredded and sub-sampled audio could not be easily reconstructed, making it difficult for an attacker to sniff any sensitive information.

The main contributions of this paper are:

- **Two methods to preserve audio privacy:** We address the concern of privacy guarantees that may be undermined by malicious software intending to sniff information from raw sounds, by preprocessing raw sounds with sound shredding and subsampling.
- **Experiments and evaluation of proposed methods:** The goal of the two proposed methods is to preserve the user's privacy without significantly decreasing context recognition accuracy. Therefore, we

HotMobile'15, February 12–13, 2015, Santa Fe, New Mexico, USA.

http://dx.doi.org/10.1145/2699343.2699366.

present the results of context recognition accuracy as well as gender and speaker recognition accuracy using shredded and subsampled sound in section 5. In addition, we also propose a sound reconstruction model using frequency content of shredded audio and quantify our findings in this section.

The rest of the paper is organized as follows. We discuss related work in section 2. Then We define a threat model in section 3 that describes the possible attacks against users' privacy. Our sound shredding and sub-sampling methods are described in section 4. In section 5, the experiments and results are elaborated and evaluated. We conclude our work in section 6.

2. RELATED WORK

Sound sensing has been shown to be useful in many context aware applications. Eronen, A.J. [6] demonstrated the usefulness of audio in recognizing environment around a device. Similarly Chu [5] used environmental sounds for the understanding context. SensOrchestra [4] achieved 87.7% recognition accuracy in determining location using audio and image. In addition to context recognition, sound can also be used for other informations. For instance, StressSense [11] used human voice recorded by smartphones to recognize stress. These experiments demonstrate the usefulness of acoustic features.

Although there is a plenty of research on using audio sensing, not much has been done on securing the collected audio data. Klasnja [9] through his work on privacy, shows strong aversion to audio sensing. He mentions "Reactions to the raw audio were nearly unanimously negative. Only two of the 24 participants (8.3 %) would consider a microphone that continuously recorded raw audio". Unfortunately, not many sound sensing applications take the privacy implications into account, therefore introducing potential attacks against user privacy.

One way to improve privacy is by extracting audio features and discarding raw audio, though now it is generally accepted that MFCC are poor features for maintaining privacy because they reveal speech [10]. The PCA of audio spectrogram is proposed to detect non-speech sounds and prevent speech reconstruction intelligently [3, 10] using filters to omit the audio. In addition, there are encryption techniques available to secure audio data, e.g. audio features encrypted by LSH key is devised to hide speech while providing cues for prosody and recognition of conversations [14]. All above methods though suitable on server, cannot be used on mobile phones. The limitations on mobile phones demand a technique, which could easily be implemented and does not consume much power, even if the application runs continuously.

3. THREAT MODEL

Sensitive information is often communicated verbally because audio is generally more ephemeral than an email or a SMS text message. However, emails and text messages are often encrypted by the applications that store them, which is rarely the case with audio data collected by sound-sensing applications. Because many sound-sensing applications collect data continuously, the audio data could reveal sensitive information.

In this paper, we are concerned with securing audio data from attackers and malicious software. To provide a clear outline of the threat model, we identify three roles involved: a user, a mobile sensing application and an attacker. A user allows a mobile sensing application to use microphone for collecting contextual information. The application continuously records audio, stores it on the phone and later uses it for context recognition tasks. The application is supposed to provide privacy guarantees to the user, but often makes no attempt to encrypt the audio data. We assume that an attacker can then get an access to the unencrypted audio files containing sensitive information (e.g., by physically stealing the device or by tricking a user to install seemingly-benign app, which will search for unencrypted audio files on the device).

To achieve the goal of privacy preserved sensing, we propose that the operating system preprocesses the audio using sound shredding, sound subsampling or both, before forwarding it to the context sensing application (as shown in Figure 1). Note that we assume that the operating system is trusted, but that the apps with access to audio data are not. For the purposes of this preliminary work, we additionally assume that an attacker has access to a limited corpus of short audio clips and cannot gain additional sensitive information about one clip from another.

4. METHODOLOGY

In this section, we introduce the technique of context recognition using audio data. Then we propose two ways to improve users privacy while collecting audio data, namely "Sound Sub-sampling" and "Sound Shredding".

The architecture of system involves mobile and server. Audio snippets are obtained from mobile OS, which after shredding and sub-sampling are stored on local memory. The stored data is later sent to server for analysis.

4.1 Context Recognition using Audio data

We define context as background environment 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. The process of context identification involves data collection, features extraction and context recognition using machine learning as discussed below:

Audio data collection: Audio data could be collected using any device with a microphone. For our experiment, a total of 35 sounds samples were recorded with a sampling rate of 8KHz sampled at 16 bit using a Nexus 4 phone.

Features extraction: First, the audio data is framed using a sliding window with window size of 30 ms, and for each of these audio frames, Mel-frequency cepstral coefficients (MFCC) of 12 vector length are calculated.

Context Recognition: Our experiment uses two machine learning algorithms namely K Nearest Neighbor (KNN) and Support Vector Machine for context classification using MFCC features.

4.2 Sound Subsampling

Identifying speech from an audio source requires a fairly continuous data but that is not the case with context recognition, which 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.

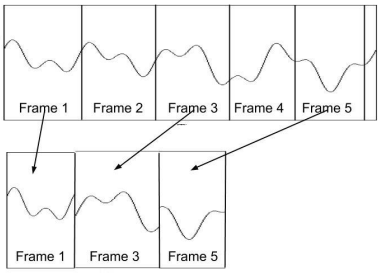

Figure 2: Sound Sub-sampling at the rate of 50%

We define Sub-sampling as the process of collecting only a part of the raw data e.g. a subsampling at 20% means only 20% of audio data is stored, i.e. only two frames out of ten audio frames are stored and rest are discarded. Figure 2 demonstrates the process of sub-sampling at 50% where every second frames is dropped during audio data collection.

4.3 Sound Shredding

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 information. 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 20-40 ms duration. These features do not change even if the sound frames are randomized as long as the frames are not changed internally.

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 3 shows the process of sound shredding. Shredded audio becomes difficult to reconstruct and replay as later demonstrated in the experiments section 5.

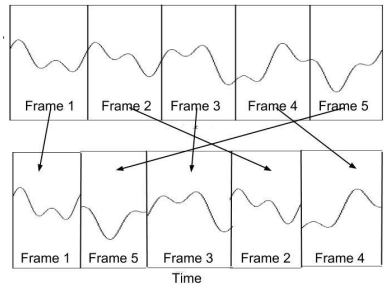

Figure 3: Sound shredding

Figure 4 shows the data collected by shredding. As the data is randomized during collection, the shredded data looks very different from sub-sampled data.

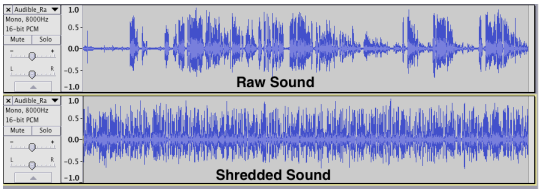

Figure 4: Sound Shredding: Raw data and shredded data

4.4 Sound Shredding and Sound Sub-sampling

In some cases of context recognition, sound shredding and sound sub-sampling can be combined for improved privacy.

5. EXPERIMENTS AND RESULTS

In this section, we describe our experiments that are divided in four parts. First we conducted experiments to determine the effect of sound shredding and sound sub-sampling on context accuracy. Then we conducted a user study to find changes in user privacy by replay of privacy preserved audio. Next we used a speech recognition engine to determine gender and speaker identification accuracy. At the end we designed and evaluated a speech reconstruction model based on frequency content of shredded audio.

5.1 Context Recognition

Audio data for the experiments was collected using a Nexus 4 phone by reading its microphone at 8000 HZ using single audio channel. In total we collected thirty-five sound samples in different contexts including: student faculty meeting, friends talking during lunch, walking, brewing coffee in cafeteria, students talking in a meeting, classroom, guest talk, laboratory etc. The experimenter used the context as the label for the audio. For each of the contexts, three sound snippets of approximately 2 minutes duration were recorded. 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.

The experiments were run with varying degree of sub-sampling (10% to 100%).

Figure 5 shows the trend of changes in the accuracy of SVN and KNN algorithms with changes 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 steep decrease in context recognition accuracy.

In addition, the experiments were also run with sub-sampled shredded sound with varying degree of sub-sampling (10% to 100%).

Figure 6 shows the trend of change in accuracy 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

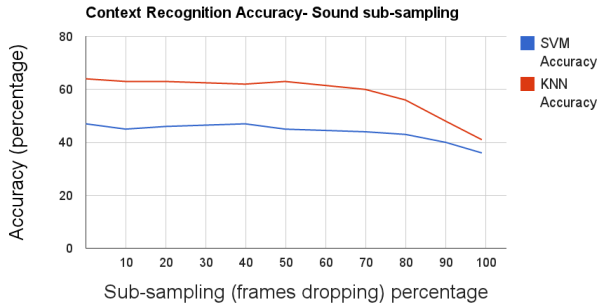

Figure 5: Context Recognition Accuracy vs. Sound sub-sampling percentage

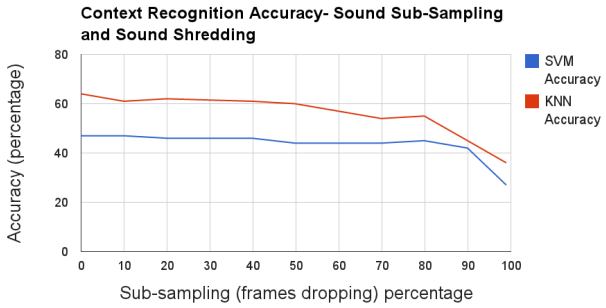

Figure 6: Context Recognition Accuracy vs. Sound sub-sampling percentage for shredded sound

frames dropping) till the sub-sampling percentage of 80%. But after around 80% sub-sampling, there is a steep decrease in the context recognition accuracy.

The above experiments and results indicate 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 privacy is discussed next.

5.2 Privacy-preservation User Study

The user study involved playing different sounds (shredded and sub-sampled) in front of users. As they hear the sound, they rated the sound on speech recognition, recognition of count of people in conversation and gender identification. Parameters and scale used for user study:

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

The scale used was 1-5, where 1 meant "Not at all" and 5 meant "Yes, I can". Over all, 10 students took the survey and the responses were averaged to use in the graph.

The data obtained was aggregated in a chart format shown in Figure 7. As it can be observed the speech recognition, one of the major concern is user privacy drastically 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.

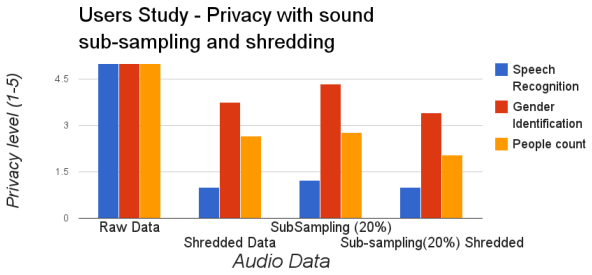

Figure 7: The user study results indicate that sound shredding can effectively protect user privacy. The speech recognition rates decrease significantly by using our approaches. The scale used was 1-5, where 1 meant "Not at all" and 5 meant "Yes, I can"

5.3 Computer-based Recognition

In the previous experiment, we studied how well can people recognize the gender, the identity and the speech given shredded and subsampled audio signal. In the following, we evaluate computer-based recognition techniques using a similar criteria. This simulates the situation of having an attacker, who gets an unauthorized access to the audio files.

We use 330 speech snippets with an average duration of 10 seconds collected from 8 users (4 male, 4 female) [8]. For gender and speaker identification we use the open source LIUM toolkit [12], which has pre-trained gender recognition model. To train the speaker identification model for our evaluation, we use one audio snippet for each user in the dataset.

As illustrated in Figure 8, subsampling and shredding does not have a significant effect on both gender and speaker identification. This confirms the results of the user study.

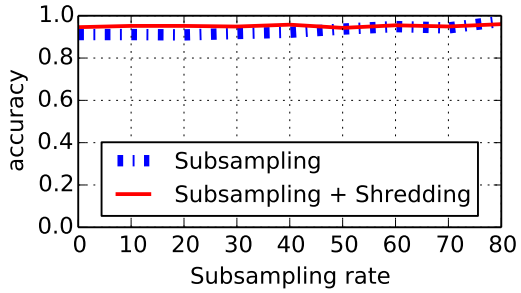

Figure 8: Subsampling and shreadding does not have a significant effect on gender prediction.

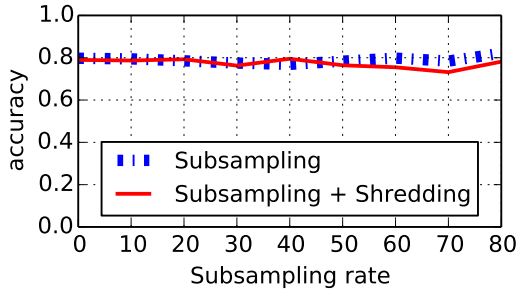

Figure 9: Subsampling and shreadding does not have a significant effect on speaker identification.

To recognize the speech we use the speech recognition system presented in Kim et al. [8]. The performance of speech

recognition is measured in Word Error Rate (WER), where the smaller WER, the more content is recovered. The WER for the original signal is 5.70%. As shown in Figure 10, with low subsampling rate, one can recover the speech relatively well. However, if an audio signal is shredded, no speech information can be recovered.

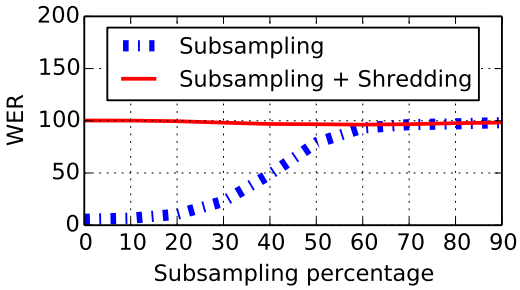

Figure 10: Speech content cannot be recovered if the audio is shredded or subsampled with high rate.

From the presented results we can observe that through shredding and subsampling the sound signals do not loose information about the gender and user identity. However, no speech content can be recovered if the audio is shredded or subsampled with a high rate. This property is highly desirable in many applications such as social life-logging. These applications aim to measure how much social interaction did a user have and how many people she met over the day without needing to know the content of user's conversations.

5.4 Speech Reconstruction Attacks

In shredded audio, all components of the original audio are present and so it is theoretically possible to reconstruct the sound, though it may be infeasible to do so because of computational challenges. As in the case of paper shredding challenge by Darpa [7], there are no single known efficient solution to reconstruct shredded sound. Possible solutions would involve a combination of approaches. Here we describe two possibilities.

5.4.1 Brute force attack

If we take a small sound sample of 10 seconds and frame width of 15 ms, there are apprximately 667 frames in the sound sample. There are n! different ways of arranging n distinct objects into a sequence, so these 667 frames can be rearranged in 667! ways. O(n!) calculations are computationally very expensive e.g. 100! is approximately 9.332622e+157, which indicates that our computer could easily run out of processing capability. Also, we need to consider the cost of analyzing the audio of each of the arrangements to get the text back, which can be either done manually or by using a speech processing system, and incurs additional cost.

5.4.2 Reconstruction based on frequency content

Shredded audio contains all frequencies present in the original audio, but they loose their original order because of shredding. The diagram 11 compares spectogram of original audio and shredded audio, where original audio was shredded in 12 pieces, looking at which one gets an impression that it could be possible to rearrange the frames back. We designed an experiment to do the same and tried a greedy algorithm approach to reconstruct the audio.

Assume we have original signal O, the shredded signal S and the reconstructed signal R. We compute d(O, S) and

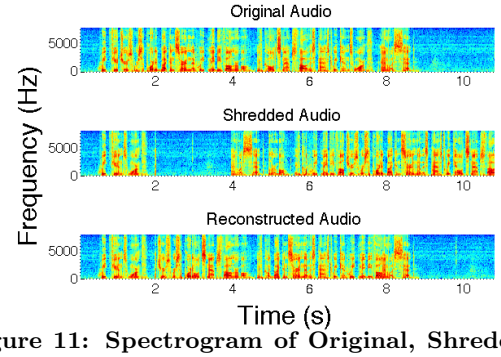

Figure 11: Spectrogram of Original, Shredded and Reconstructed audio. The original audio was split in 12 pieces and shredded. In this case, the greedy algorithm could partially reconstruct the original audio. As the no of divisions (split) increases, it becomes increasingly difficult to reconstruct the original audio as shown in figure 12.

d(O, R), where d() is Euclidian distance function comparing two audio encodings. We define audio encodings as the arrangement of audio frames e.g. an audio signal O can be represented as a string abcdef.... where each character represents an audio segment, whereas a shredded audio S will be represented as dbmkc.... comprising entirely of characters present in audio O, but in a random order. We compute d(O,S) as the Euclidian distance between strings abcdef.... and dbmkc...., where each character represents a number e.g. a=1, b=2 etc. In addition, we also calculate similarity between O and R using Longest common subsequence (LCS) algorithm which uses abcdef... and dbmkc... as two strings obtained from the method described above.

To reconstruct signal O from signal S, we take inspiration from paper shredding experiment [7] where right edge of a shredded part matches the left edge of the shredded part on the right. Similarly, for audio spectrogram, the frequencies present on the right most window of a segment will be similar to frequencies present in the left edge of the segment on the right. Based on this idea, we compute spectrogram of segments of shredded audio, which gives amplitude and phase of all frequencies present in smaller segments d,b,m,k.... Then we start with the first frame of shredded signal S, namely d and use greedy approach to search the remaining frames in S, to find the closest match for frequency amplitude present in the rightmost window of segment d. If closest match of d gives k, then we construct signal as "dk" and then start another greedy search for k. This way we can reconstruct complete audio R like "dkmabc....". We then compute d(O,R), the Euclidian distance comparing O and R. The measure of d(O,R) gives us how much successful we are in reconstructing the shredded audio. In addition to the Euclidian distance we also calculated similarity using Longest common subsequence algorithm.

The results shown in figure 12 reveals that the thinner the shredding is, the more difficult it is to reconstruct the audio. The audio signal which was divided in 5 or lesser segments, it was possible to reconstruct the audio, but as the number of divisions increase (shredding thickness decreases), it becomes increasingly difficult to reproduce the original audio based on frequency content.

6. CONCLUSION AND FUTURE WORK

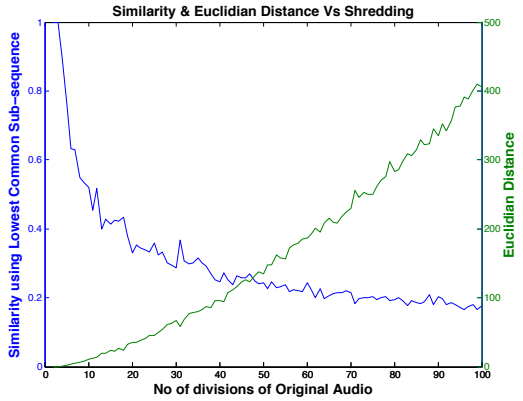

Figure 12: Euclidian distance and Similarity (using longest common sub-sequence) between Original and Reconstructed audio Vs No of divisions of Original Audio. The result indicates that the thinner the shredding, the more difficult it becomes to reconstruct the original audio.

Audio is a valuable source of contextual information, which is crucial for many context-aware mobile applications. However, beside context information the captured audio signals often contain sensitive speech content. In this work, we show that sound shredding and subsampling are effective means for making speech not recognizable, while preserving sufficient information for context, gender and speaker recognition. Through the experiments, we showed that no speech content could be recognized from the processed signal by either human or automated computer techniques.

In future work, further studies are needed to understand the effectiveness and robustness of the proposed approaches. Since both sound shredding and subsampling are meant to be run directly on mobile devices, additional experiments are needed to analyze the battery consumption and the computational complexity of these approaches. Although we provided a theoretical analysis of attacks aiming at reconstructing the original signal, more sophisticated attacks needs to be explored to study the effectiveness of the proposed approaches.

7. ACKNOWLEDGEMENT

This research is supported in part by the National Science Foundation under the award of 1346066: SCH: INT: Collaborative Research: FITTLE+: Theory and Models for Smartphone Ecological Momentary Intervention.

We would like to thank all the reviewers for their insightful comments and suggestions which have greatly helped us to improve our work.

8. 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] F. Chen, J. Adcock, and S. Krishnagiri. Audio privacy: reducing speech intelligibility while preserving environmental sounds. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 733–736. ACM, 2008.

[4] 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*, volume 76 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 195–210. Springer Berlin Heidelberg, 2012.

[5] S. Chu, S. Narayanan, and C.-C. J. Kuo. Environmental sound recognition with time-frequency audio features. *Trans. Audio, Speech and Lang. Proc.*, 17(6):1142–1158, Aug. 2009.

[6] 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, Jan 2006.

[7] T. Geller. Darpa shredder challenge solved. *Commun. ACM*, 55(8):16–17, Aug. 2012.

[8] J. Kim and I. Lane. Accelerating large vocabulary continuous speech recognition on heterogeneous cpu-gpu platforms. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3291–3295. IEEE, 2014.

[9] P. Klasnja, S. Consolvo, T. Choudhury, R. Beckwith, and J. Hightower. Exploring privacy concerns about personal sensing. In *Proceedings of the 7th International Conference on Pervasive Computing*, Pervasive '09, pages 176–183, Berlin, Heidelberg, 2009. Springer-Verlag.

[10] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel. Accurate and privacy preserving cough sensing using a low-cost microphone. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 375–384. ACM, 2011.

[11] 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.

[12] S. Meignier and T. Merlin. Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*, volume 2010, 2010.

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

[14] D. Wyatt, T. Choudhury, and J. Bilmes. Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *INTERSPEECH*, pages 586–589, 2007.