Characterizing the Effect of Audio Degradation on Privacy Perception And Inference Performance in Audio-Based Human Activity Recognition

Dawei Liang
The University of Texas at Austin
Austin, USA
dawei.liang@utexas.edu

Wenting Song
The University of Texas at Austin
Austin, USA
wentingsong@utexas.edu

Edison Thomaz
The University of Texas at Austin
Austin, USA
ethomaz@utexas.edu

ABSTRACT

Audio has been increasingly adopted as a sensing modality in a variety of human-centered mobile applications and in smart assistants in the home. Although acoustic features can capture complex semantic information about human activities and context, continuous audio recording often poses significant privacy concerns. An intuitive way to reduce privacy concerns is to degrade audio quality such that speech and other relevant acoustic markers become unintelligible, but this often comes at the cost of activity recognition performance. In this paper, we employ a mixed-methods approach to characterize this balance. We first conduct an online survey with 266 participants to capture their perception of privacy qualitatively and quantitatively with degraded audio. Given our findings that privacy concerns can be significantly reduced at high levels of audio degradation, we then investigate how intentional degradation of audio frames can affect the recognition results of the target classes while maintaining effective privacy mitigation. Our results indicate that degradation of audio frames can leave minimal effects for audio recognition using frame-level features. Furthermore, degradation of audio frames can hurt the performance to some extend for audio recognition using segment-level features, though the usage of such features may still yield superior recognition performance. Given the different requirements on privacy mitigation and recognition performance for different sensing purposes, such trade-offs need to be balanced in actual implementations.

CCS CONCEPTS

• **Security and privacy** → **Domain-specific security and privacy architectures**; **Privacy protections**; • **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**.

KEYWORDS

Privacy; Mobile Sensing; Audio Processing; Activity Recognition

ACM Reference Format:
Dawei Liang, Wenting Song, and Edison Thomaz. 2020. Characterizing the Effect of Audio Degradation on Privacy Perception And Inference Performance in Audio-Based Human Activity Recognition. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20), October 5–8, 2020, Oldenburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3379503.3403551

1 INTRODUCTION

With development of personal audio sensors such as smart phones and wearable computers, audio-based sensing and recognition have been shown to be of value for a wide range of applications, such as health monitoring and personal assistance. Compared to traditional inertial features, audio provides rich contextual information about users and the environments. Consequently, audio has been explored over the last decade in a variety of human-centered recognition efforts. For instance, Eronen et al. [10] proposed a pilot study to detect background environment (context) based on sound with statistical learning methods. Lu et al. [25] developed a mobile system to classify human activities at different levels by using audio data collected from users. Sounds can also be used to infer affect and physiological markers. For example, Nasir at el. [26] used sound recordings from couple therapy to analyze emotion in conversations.

Most current studies of acoustic sensing and recognition are realized with mobile or wearable sensors. In the context of real-world activity recognition, for example, the sensing devices may not be able to anticipate the duration of the target activities. Hence, to enable reliable detection of sound events in real time, sensing devices must be capturing and recording audio data continuously, which raises privacy concerns [16]. Also, by capturing speech, sound is pervasive and typically mixed with information from multiple sources, making it difficult for users to know what information has been captured in their recordings. Consequently, the privacy concerns due to unexpected capture by the usage of personal sound sensors largely make the public alert to the generalization of such technology in their daily life.

To reduce privacy concerns in personal sound-based applications, many prior works have been developed. For example, audio can be distorted [3], segmented [39] or partially obfuscated [22]. Given the computational limits, Kumar et al. [19] explored a simple audio frame degradation method by random frame dropping. It can be a generalizable and preferable choice for real-world sensing platforms. It would therefore be interesting to study whether and how the leverage of intentional audio degradation can help the mitigation of people's privacy concerns in real sensing settings. Recently studied

MobileHCI '20, October 5–8, 2020, Oldenburg, Germany

ACM ISBN 978-1-4503-7516-0/20/10…$15.00
https://doi.org/10.1145/3379503.3403551

by Perez et al. [31], human perception of privacy is subjective in real-world scenarios and can change with the contexts being recorded and amount of information that people perceive from the sounds. Hence, it would also be insightful to incorporate the actual acoustic contexts when evaluating the effectiveness of the proposed privacy protection methods.

In this paper, we first qualitatively and quantitatively study how people's privacy concerns due to the exposure to personal acoustic sensing devices can be mitigated with intentional audio degradation based on a user study with 266 participants and 4 sensing scenarios. Our findings suggest that the perception of privacy can be significantly reduced with high levels of audio degradation. We then conduct generalization analysis to determine how intentional degradation of audio frames can affect audio-based activity and context recognition performance while achieving the promising privacy mitigating effects. By leveraging real-world and online sound data with common frame-based and segment-based audio features, we determine that frame-level audio degradation methods can be effective for the mitigation of people's privacy concerns while leaving minimal effects on activity and context recognition based on audio frame features. Furthermore, we show that it is still possible to achieve promising performance for the recognition of segment-based spectrogram features after degradation of audio frames, but the performance can drop depending on the degradation and classification choices. The trade-offs of such choices need to be balanced in actual implementations.

2 RELATED WORK

2.1 Audio-Based Activity and Context Recognition

Sounds can be used to capture syntactic features of real-world human activities, and therefore are widely used for human activity recognition. Prior research has shown such feasibility on different sensing platforms and with different application domains. Early work by Eronen et al. [10] proposed the usage of real-world sound to recognize common contexts. To advance elderly care, Chen et al. [4] provided an audio-based solution for the detection of 6 bathroom-related activities. In recent years, sound-based activity recognition has been studied in a more human-centered setting. For example, Yatani and Truong [40] developed the *BodyScope* system that could be used to detect 12 human activities related to throat movement. Thomaz at al. [36] proposed the inference of eating moments based on statistical sound features from a wrist-mounted sensor. Lu et al. [25] and Rossi et al. [34] studied the recognition of various human activity classes using statistical classifiers on mobile platforms. With deep learning, Lane et al. [20] developed the *DeepEar* as a pilot mobile application using deep learning for multi-task sound-based detection. Becker et al. [2] developed the *GestEar* for gesture recognition with neural nets on a smartwatch. Laput et al. [21] proposed a plug-and-play activity recognition system leveraging sound features from multiple online data sets. Similarly, Liang and Thomaz [23] explored the usage of large-scale acoustic embedding features from public YouTube video sound clips to empower activity recognition in the wild.

2.2 Privacy Concerns in Personal Sound Sensing

Sound signals are pervasive in the environment, usually consisting of compound information from various sound sources. In unobtrusive sensing, the subjects are typically unaware of the sensing process. Such factors lead to unexpected exposure of the users to the audio recordings while leveraging sound-based sensing applications. As presented by several prior work [6, 16, 17, 29], the unique mapping between sound and certain human activities or locations can increase the risks of revealing user's private information. Especially studied by Klasnja et al. [16], people are significantly more concerned about continuous sound sensing in their daily life comparing to some other types of sensing modalities for activity recognition such as inertial features or GPS signals. Perez at al. [31] further quantified how people's privacy concerns towards sound recordings can be affected by the actual recorded contexts.

As a natural way of how people interact with each other and with the environment, speech has always been considered to contain user's personal information [17, 29]. As pointed out by Raij et al. [33], the public tends to be more worried about the privacy risks regarding their conversations captured by wearable and mobile sensors than other types of behavioral measurement. Diao et al. [8] discussed the threats from acoustic sensors such as voice assistants. Ammari et al. [1] investigated the types and cause of people's privacy concerns with smart voice assistants. The concerns can be raised especially when users are not aware of the recording process and have no access to the data being shared. Correspondingly, many efforts have addressed privacy protection regarding human speech in sound sensing.

2.3 Privacy Protection for User Audio

One common way of speech projection is to detect human conversations in the recorded audio and filter out the intelligible information. Wyatt et al. [39] proposed the idea of using clustering methods and pairwise distribution of the speakers to detect human speech in audio. However, this largely restrict the types of features available for sound event recognition. Alternatively, Liaqat et al. [24] proposed the usage of Linear Predictive Coding (LPC) coefficients to detect human speech that could be filtered out later. Chen et al. [3] also showed that speech intelligibility can be significantly reduced if vowels of the speakers are altered. This is useful in cases where a set of human vocalics can be accessible. With the development of deep learning approaches, Vatanparvar et al. [38] used a Generative Adversarial Network (GAN) architecture to generate artificial speech and replaced the original data. A more recent study by Nelus et al. [28] added stochastic feature representation for a neural network-based feature extractor to alleviate speaker identification. While these methods are effective for the mitigation of user privacy risks, they also lead to considerable computing burdens with extra processes of model training and feature transformation.

In addition, prior work [22, 41] explored the simplification of audio spectrogram for privacy protection. Yet this is limited to specific types of acoustic features and still requires extra computation efforts that can be obstacle on personal sensing devices. As an improvement, Kumar et al. [19] blurred the sensed audio simply by degrading the audio frames. This is promising for light-weight

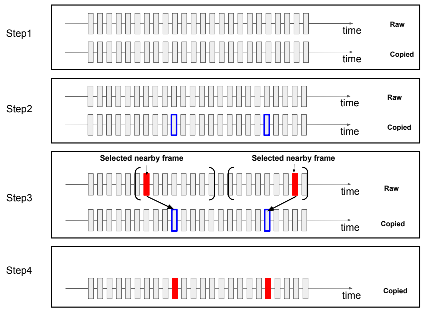

Figure 1: Visualization of the replacement-sampling process with sample frames. (Step 1: The raw audio is copied; Step 2: Target frames to be replaced are selected in the copied audio clip; Step 3: Each target frame is replaced with a frame randomly selected from its corresponding neighbouring pool in the raw clip; Step 4: The raw audio clip is deleted.)

smart phone or wearable sensing. To the best of our knowledge, however, very few prior works attempted to study the change of people's privacy concerns in audio-based recognition tasks with the actual implementation of such audio degradation methods. Furthermore, it would be insightful to characterize the balance between effective privacy protection and the cost of performance in audio recognition.

3 EVALUATION OF PRIVACY CONCERNS WITH AUDIO DEGRADATION

Audio signals are widely used for human-centered recognition tasks. Audio sensors integrated in personal wearable and mobile devices can not only catch the target context and activity sounds, but also speech information of the users and bystanders. Prior work [8, 33] already showed the corresponding privacy risks and people's concerns towards the unexpected capture of their intelligible speech information. Hence, in this section we quantify how such concerns can be mitigated with intentional audio degradation.

3.1 Audio Degradation

To degrade the recorded audio, the prior work by Kumar et al. [19] proposed frame dropping (down-sampling) and order shuffling methods. Conversely, the frames can also be randomly up-sampled. We also explored an alternative replacement-sampling strategy to keep the temporal size of the sound clips consistent for all test cases. The basic idea is to randomly replace some of the original sound frames with their nearby frames in the audio. All these methods follow the same principle that speech can be degraded without affecting much the recognition of the target sounds as long as we disrupt the global audio sequence while maintaining the acoustic patterns within a frame unmodified.

Figure 1 shows the general process of audio degradation with replacement-sampling. Given the original audio clip, a proportion δ of target frames to be processed are randomly selected out of the sound sequence without replacement. For each of them, a neighbouring frame within a pool range of K is randomly selected and used to replace the corresponding target frame. The process repeats until all target frames are replaced. To avoid the effects of accumulation during frame replacement, the neighbouring frames are selected from the original audio clip and the replacing process is implemented on a copied version of the clip. We controlled the audio degradation levels by varying the values of δ.

3.2 Online Survey

We then conducted a survey on the Amazon Mechanical Turk online platform to evaluate people's privacy concerns with the degraded audio. The general idea of the study is to present the participants with speech corpus together with simulated sensing scenarios. The corpus may or may not be degraded and people's privacy concerns were studied by recording their attitudes of being captured in a similar way as presented. As described by Dimiccoli et al. [9] and Perez et al. [31], human's sense of privacy can change with different contextual cues. Hence, in the study we incorporated such factors by applying 4 simulated sensing scenarios (two outdoor cases and two indoor cases): 1) couple conversation in the home; 2) family dinner; 3) phone chat at a public space; 4) interaction with friends at a party. They also roughly match the common types of interpersonal relationships (family members, friends, acquaintances) as described by Granovetter et al. [13] with common backgrounds in daily living (private and public space). The complete scenarios were given as follows:

- *You are at a park and having a phone conversation with a friend. Someone else nearby is wearing a device (e.g., a smart watch) that is continuously capturing ambient sounds, which includes your phone conversation.*
- *You are in a restaurant at a party and chatting with others. Someone else nearby is wearing a device (e.g., a smart watch) that is continuously capturing ambient sounds, which include your conversation.*
- *You are at home and talking to your husband/wife. A device placed in a fixed location in your home is continuously recording audio of the environment, including your conversations.*
- *You are at home and having dinner with your family. A device placed in a fixed location in the dining room is continuously recording audio of the environment, including your conversation.*

For each of the scenarios, we prepared featured conversation clips varying from 15-18 seconds (around 35 to 45 words) in length. The dinner corpus consists of simple conversations between two family members on the food. The party and the couple corpus consist of chat between a couple about an appointment and two friends on a party respectively. The phone corpus was a short monologue by a male calling for an absence of his daughter. All corpora were generated with a normal speed by voluntary native English speakers. To simulate the actual sensing environment, the volunteers were also asked to perform the speech as if they were in the context and for the party corpus, in particular, the sound was recorded in an actual restaurant with background noise. A pilot lab study was conducted to ensured that the audio recordings could be perceived clearly and correctly by human listeners.

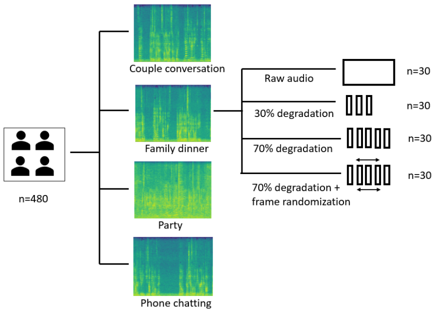

Figure 2: Division of participants for the Amazon Mechanical Turk online survey. The subjects are randomly divided into 4 sensing scenarios with 4 audio degradation levels (including the control group). At least 30 survey requests were sent for each sub-group and participants of the same sub-group were presented with the same scenario and corpus.

All the above frame re-sampling methods are effective to degrade the audio, but we applied only the replacement-sampling strategy so that the size of corpora can be consistent across the tested cases. We determined three levels of audio degradation: low (δ=30%), high (δ=70%) and full (δ=70% with frame order randomization). The neighboring range K was selected as 50 each side. During processing, the audio was sampled at 16kHz mono and framed for every 60 ms without overlaps. We chose the degradation values since the audio was only slightly degraded when δ=30%. When δ=70%, the speech was significantly degraded and the intelligible information could not be recognized by human listeners in most cases. The intelligible information no longer remained when frame order randomization was applied. There was also a control group in each sensing scenario where the audio was simply left as it was. Hence, there were 16 speech corpora in total.

Figure 2 shows the overall process of the survey. Each sensing scenario was divided into 4 sub-groups with at least 30 survey requests sent each. Before the survey started, background information was presented with the overall goal and process of the survey. The participants were encouraged to use their headphone while performing the study. Since participants of different age and technological backgrounds may have different sense of privacy towards new technologies, the study began with general questions of the participants' age and a statement " *I consider myself to be a technology-savvy person*". The subjects could respond to the statement based on 5-point Likert-scale with options: *Strongly agree*, *Agree*, *Neither agree nor disagree*, *Disagree* and *Strongly disagree*. The participants were then presented with one of the sensing scenarios and one of the speech corpus. Participants of the same sub-group were presented with the same scenario and corpus. Furthermore, they were not told if the corpus had been degraded.

The study was then conducted in two phases. The first part aimed to confirm that the degradation did alleviate the intelligible information from the participants' perspectives. We asked the participants to transcribe the speech corpus in a text box. They were encouraged to guess if they failed to recognize any of the words. Besides, they were allowed to replay the audio as many times as needed. It was also followed by a 5-point Likert-scale regarding the statement:"*How confident are you that your transcription is accurate?*". The available options were: *Very confident*, *Confident*, *Neither confident nor unconfident*, *Not confident* and *Not confident at all*. The word error rates (WER) of the transcripts were calculated to quantify the information accurately captured by the listeners. We leveraged the Python Jiwer package [37] based on the Wagner-Fischer algorithm [27] with removal of some abbreviations and punctuation in the sentences.

In the second part of the survey, we then asked the participants to consider the statement: "*I would not mind being captured in an audio recording like this.*" They could response with a 5-point Likert-scale of options: *Strongly agree*, *Agree*, *Neither agree nor disagree*, *Disagree* and *Strongly disagree*. To further obtain insight of their selections, we also added a text box for text input and asked the participants to briefly explain their selections. The goal of the designed questions is to determine if audio degradation can be effective to mitigate people's privacy concerns if they are captured in personal sound sensing, and if so, how such mitigation of concerns can be quantified. We hypothesized that **the intentional degradation of audio can mitigate people's concerns of being captured by personal sound sensing in the given sensing scenarios.** We studied the results by comparing the proportion of people choosing *Strongly disagree* and *Disagree* with a two-proportion z-test.

Participants were from the global pool of the Mechanical Turk system and we did not require specific skills for the study subjects. However, we required that all participants must be MTurk masters to ensure the quality of the response. The time limit was set as 10 minutes for each survey. After the study, the participants were compensated with a range of 0.25-0.5 US dollars. The study was also approved by an IRB protocol before implemented.

3.3 Results and Findings

The survey lasted for around 3 weeks. Before quantifying the responses, we noticed that there were high proportion of outliers in some of the groups. As discussed by Crump et al. [7], the outliers are expected on the MTurk platform since very little environmental control can be applied to the survey participants. The number of outliers can depend on several potential factors include difficulties and duration of the tasks, compensations and specific requirements of the devices needed. In our study, two types of outliers were observed. One was repetition of workers in the same sub-group such as responses with the same worker IDs or text inputs of unreasonable repetitions. The other was entirely irrelevant responses regarding the contexts such as an input of "*I like the speech*" as the explanation of choosing a privacy level. Hence, a filtering process was conducted by two researchers of the paper to independently determine the potential outliers. The response would be discarded if there was agreement between the two decisions. After the filtering process, we were able to obtain 266 valid responses in total.

Among the survey participants, we found that the top-two age groups were 20-30 and 30-40 with the proportion of 44.2% and 22.1% respectively. Besides, 82.3% of the participants chose *Strongly agree* or *Agree* regarding the statement that they were technology-savvy.

Degradation	WER	Confidence
None	34%	93%
δ=30%	52%	54%
δ=70%	93%	27%
δ=70% + Randomization	96%	21%

Table 1: Mean word error rates (WER) and proportion of participants feeling *Very confident* / *Confident* regarding the statement:"*How confident are you that your transcription is accurate?*". The speech corpora become almost unintelligible when the audio frames are highly degraded. (Lab test WER: 6%, confidence level: 100%)

We started by quantifying people's perception of the intelligible speech information with and without the audio degradation. Before calculating the WER, the transcripts were re-organized with consistent format to avoid input error. Table 1 presents how the averaged WER of the transcripts for the four sensing scenarios and participants' confidence levels of their transcription change with the audio quality. We noticed that the WER for some of the participants could be high even when no degradation was applied mainly due to typos and abbreviations, and we kept the original versions for the fidelity of the responses. As expected, there is a clear negative relationship between the reliability of the transcripts and levels of audio degradation. When the audio was severally degraded ($\delta \geq 70\%$), the WER already reached to over 90%, meaning that the speech was almost unintelligible in such cases. Besides, we noticed that it was confusing for some of the participants to report their confidence levels if they had already failed to respond to the transcription section, so we counted the confidence levels only over the population whose WER of the transcript was less than 95%, i.e., at least correctness of around 3 words. In table 1, the ratio between participants in each degradation group reporting *Very confident* / *Confident* and the whole population of that group was reported. Unlike the rise of the WER, people became less confident towards their transcription at higher levels of audio degradation. The audio with over 70% degradation was already unintelligible. By combining both the WER and confidence results, we can see that **degradation of the audio frames at a strong level effectively removes the intelligibility of speech information for the given sensing scenarios.**

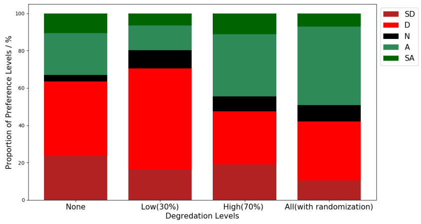

Figure 3: Distribution of survey responses to the statement "*I do not mind being captured by audio like this*" with four sound quality levels. There is significant drop (p<0.05) of people's privacy concerns of being captured by the sensed audio when comparing the degraded groups of δ = 70% or full degradation with the control groups. (SD: Strongly disagree, D: Disagree, N: Neither agree nor disagree, A: Agree, SA: Strongly agree)

Scenarios	$\delta \leq 30\%$	$\delta \geq 70\%$
Couple	68.6%	50.0% (27.1% ↓)
Dinner	60.0%	43.3% (27.8% ↓)
Party	72.7%	38.7% (46.8% ↓)
Phone	65.8%	50.0% (24.0% ↓)

Table 2: Proportion of participants choosing *Strongly disagree* and *Disagree* regarding the statement "*I do not mind being captured by audio like this*" for the four sensing scenarios. The privacy concerns drop in all test cases when the audio is strongly degraded.

We then studied how participants' privacy concerns of being captured in the audio could change with the audio degradation. Figure 3 shows the proportion of each choice regarding the statement "*I would not mind being captured in an audio recording like this.*". As we can see from figure 3, participants tended to feel less worried about being captured in the sensing scenarios when the audio was strongly degraded ($\delta \geq 70\%$). By using the two-proportion z-test, we determined that the proportion of participants choosing *Strongly disagree* and *Disagree* significantly dropped ($p<0.05$) from the mild degradation groups (no degradation or δ = 30%) to strong degradation groups (δ = 70% or with frame shuffling). The effect size between the control groups and groups of δ = 70%, (δ = 70% + randomization) is 0.32 and 0.43 respectively. The responses are further quantified for each sensing scenario in table 2. We grouped the degradation levels as mild ($\delta \leq 30\%$) and strong ($\delta \geq 70\%$), and we compared the results for the sensing scenarios individually. As we can see from table 2, the participants' concern levels drop by an amount varying from 27% to 47% in the tested scenarios after the audio was disrupted. Hence, we concluded that **intentional degradation of the audio frames at a strong level is effective for the mitigation of people's privacy concerns in the given sensing scenarios.** We then investigated the reasons why people would feel less concerned in such cases by checking the text responses. As pointed by a participant, the fact that the audio was no longer intelligible generally eased their concerns: "*No one can understand it, so what I said on it wouldn't matter.*" Similar explanations can be seen from other participants: "*It was all broken up and you couldn't tell what anyone was saying.*"; "*It's scrambled in a way that makes speech inaudible.*"

Nevertheless, some people still worried about being recorded without permission even at high levels of audio degradation:"*I don't really want to be eavesdropped on, even if it is garbled.*"; "*Despite this audio being completely masked so nothing's really understandable, it's uncomfortable to have someone that I have no relation to recording everything I do around them. It makes me nervous about what that person might use the recordings for an if there's some way to unmask them.*".

As expected, high proportion (63.5%) of the participants felt concerned about being captured by sound sensing when no audio degradation was implemented at all. We did not see significant difference among the sensing scenarios of the control groups. By checking the responses of the participants, many of them explained their concerns: "*The conversation is private and I don't want devices to record this kind of conversation. The information is slightly sensitive and I would mind if it gets recorded.*"; "*I would not like to have my private conversations recorded.*" In addition, we noticed that the speech contents could also play an important role when people judged privacy. For example, a participant who did not mind being captured in the original audio of couple conversation mentioned: "*I think that this is a normal conversation between a husband and a wife and I normally face it.*". Another participant choosing *Agree* in the groups of no degradation also stated: "*There is no sensitive information in this recording, so I am not concerned by this type of audio recording.*" In other words, people's perception of privacy towards general sound sensing could partially depend on the intelligible information being captured. Some people can feel less concerned if they can determine that there is no sensitive information in the sensed sounds.

We did not find statistical difference between the degraded groups of δ = 30% and the control groups. In fact, the proportion of the negative responses was even slightly higher when the audio was degraded at a low level. A possible reason is that people were not able to fully capture the details of the conversations as in the control groups and they could not determine if the recorded sounds contained sensitive information that was not perceived directly, especially in their opinions that the intelligible information might still be recognizable. For example, some of the participants pointed out: "*It's not totally intelligible but enough context comes through to raise suspicions among people who like to be suspicious.*"; *Even though it's hard to discern exactly what's being said, I can make out a fair bit of the conversation, and I wouldn't want others to have access to private conversations like this. It also depends on who would be accessing the recording and how it would be used. I would feel less worried about it if it were totally anonymized and if the voices were altered.*" Another possible factor was that speech slightly scrambled in such ways may distort people's perception of the speech context. For example, some participants misjudged the scenario of family dinner as couple argument: "*It sounded like an argument and I don't want it captured.*"; "*It's argumentative and puts both people, particularly the man, in a bad light.*".

4 MORE DISCUSSIONS ON AUDIO DEGRADATION APPROACHES

Audio degradation by re-sampling the audio frames is useful for real-world sensing due to its simplicity in computation and its generalization capabilities. As discussed in Section 3, audio frame degradation is valid for speech protection since the sequential information of speech can be disrupted without affecting much the acoustic patterns of the target sound frames. When the recognition is based on independent instances/frames, the recognition performance can be maintained because the acoustic features within frames remain unmodified. When implemented with temporal features (e.g. spectrogram or texture window of frames), however, the sound recognition performance may be affected depending on the consistency of the frames within a sound segment. We will discuss such effects in Section 6.

In the audio recognition tests, we will study different types of frame-level audio degradation methods. The basic approaches proposed by Kumar et al. [19] are frame order randomization and down-sampling. While randomization can entirely remove the intelligible information in the audio, it can also be susceptible to the brute-force attacks. In addition, the global sequential characteristics of the audio are broken so that it is not feasible for sequence-based audio recognition frameworks. Audio down-sampling generally helps to alleviate the problems of the brute-force attacks. However, it may bring the cost of recognition performance, especially at high degradation levels. As an inverse process, up-sampling the frames by random cloning and insertion can avoid the information loss in audio degradation.

As a potential alternative, the frame replacement-sampling strategy applied in the Mechanical Turk study has the advantage of keeping the temporal size of the audio. It is actually a mixture of both audio down-sampling and up-sampling. Rather than simply dropping the frames, the frames replaced can still appear again if selected by the replacement processes of their neighboring frames. This is the reason why we need the replacement to be implemented on a copied audio clip. Figure 4 shows a sample case where the target frames to be replaced are within the candidate pool for selection of their neighbor, and hence are possible to re-appear again. With the increase of the degradation levels, there will be higher chances that such neighboring pools can overlap with each other and all frames, including the frames replaced, may even be up-sampled for multiple times. As a result, the loss of frames can slow down at higher degradation levels. Since all such methods are feasible for audio degradation, we will compare them together in the recognition analysis.

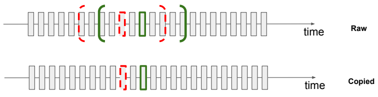

Figure 4: Sample case where the target frames are in the neighbouring pools of each other. In such cases, a replaced (dropped) frame in the copied audio clip may still be selected from the raw clip and be used to replace its neighbour, and hence up-sampled.

5 RECOGNITION PERFORMANCE ON SINGLE AUDIO FRAMES

5.1 Test with field audio

The effect on recognition performance is first examined based on single audio frames of human activity recordings. We recruited 14 participants to collect the real-world activity sounds. Our data consists of 15 activities of daily living *(Bathing/Showering, Squeezing Juice, Boiling Water, Brushing Teeth, Chatting, Chopping Food, Flushing Toilet, Frying Food, Using Microwave Oven, Listening to Music, Shaving, Outdoor Strolling, Watching TV, Floor Cleaning, Hands/Face*

Washing) with varying duration of 165 seconds to 2,175 seconds. The total size of the recordings is around 3.4 hours excluding the transitions of activities.

To collect the audio, the participants were asked to continuously perform the activities and the sound was recorded using a smart phone (Huawei P9) placed nearby. The collection was conducted in the participants' actual home environments, and the participants followed a pre-defined scripted instruction while performing the target activities. Sample instruction included *"first head to the bathroom, wash your hands and face"* or *"after juice is prepared, please warm some food using the microwave oven"*. The subjects then simply followed the instruction and an experimenter (one of the authors of the paper) would follow at a distance for notes. The proposed setting aimed to help the participants to perform the activities in their natural ways so that the recordings could reflect the actual living patterns better.

Each target activity was only performed once per subject with no timing limits. For class 'watching TV', participants were asked to watch 5 different channels for about 30 seconds each to increase the variation of the data. For activity 'listening to music' , the subjects were allowed to either play their own musical instrument (e.g. piano) or just listen to online musics chosen by themselves. Besides, the 'shaving' activity was only for male participants. The participants were encouraged to use their own equipment and the home devices (e.g., toilet fan and refrigerator) were left as natural. The participants were required to sign an IRB form before the sound collection and they were compensated 5 US dollars per person after the study.

The audio was converted as 16-bit depth mono and sampled at 16KHz. We applied a consistent 60-ms window size with no overlaps for framing. Both feature extraction and audio degradation were implemented with the same frame size. The features we used for the activity recognition task were 13 mel-frequency cepstral coefficients (MFCC) at the single frame level. We then applied a 1-dimensional convolutional neural network (CNN) as the classifier. Based on initial tuning on the user data, we determined an architecture with three convolutional layers and a fully-connected output layer. The number of channels for the convolutional layers were 16, 32, 32 respectively. The filter size was 8 with a single stride. All convolutional layers were activated by the rectified linear unit activation with the same padding. During training, we applied the stochastic gradient descent optimizer with a learning rate of 0.01. The loss function was categorical entropy loss. In addition, we added a dropout layer [35] of 0.1 for each convolutional layer. The neural network was implemented using the Python Keras [5] package.

The audio data was then split with a 5-fold basis. However, we did not shuffle the data globally at this stage to avoid the overlaps of similar sound frames between the training and evaluation sets. We then obtained the highest evaluation accuracy observed in each fold and calculated the average of the 5-fold accuracy as the overall performance metric. The performance was first reported based on the original sounds. With the same CNN, we implemented two paradigms of study. In the first study, both the training and the evaluation data was degraded. In the second study, only the training set was degraded. Since the speech mitigation levels are mostly affected by the degradation level δ, we tested the methods with only δ = 50%, 70% and 90% since the privacy mitigation is only valid at high δ levels. The goal of the study is to investigate if frame-level audio degradation can maintain the recognition performance with features extracted only at a single frame level. Also, we hope to gain deeper understanding of how different ways of adopting the audio degradation methods can affect the model generalization on the sounds.

Degrading Method	δ **= 50%**	δ **= 70%**	δ **= 90%**
Dropping (training & test)	+0.00	-1.20	-3.29
Cloning (training & test)	+0.29	+0.66	+0.25
Replacing (training & test)	-0.12	-0.37	+0.16
Dropping (training set only)	+0.16	-0.88	-2.51
Cloning (training set only)	+0.37	+1.16	+0.28
Replacing (training set only)	+0.10	-0.29	-0.01

Table 3: Change of averaged 5-fold accuracy (in %) based on the field audio frames when comparing the results of audio degradation to the results without any degradation (59.16%). The accuracy values are examined when both the training and test sets are degraded or when only the training sets are degraded. The accuracy remains generally similar with and without degradation, except that it slightly drops when the audio is severely down-sampled.

5.2 Field test results

Table 3 shows the results of two paradigms of studies: testing on the distorted sound frames and testing on the raw audio frames. In addition to the averaged 5-fold accuracy, we also calculated the variance of the 5-fold performance. In the study, the highest evaluation accuracy values were mostly observed after early stages of training (5-12 epochs). The mean 5-fold recognition accuracy based on the original sounds is 59.16% with variance of 7.21%. The results obtained using the frame order randomization method is 59.56% with a variance of 8.86%. The performance is as expected since the shuffling process does not affect the frame-level feature extraction. We see that the audio recognition performance remains stable around the original results, indicating that the implementation of the degrading methods generally does not harm the classification. However, we notice that down-sampling the frames too much can still lead to decrements on the accuracy, as also reported in the study by Kumar et al. [19].

5.3 Test with online public audio

To further demonstrate the effects, we also conducted audio frame recognition by leveraging the public ESC-50 audio dataset [32] which consists of 5-second audio clips of common nature and behaviour classes observed in daily life. These sounds were extracted from the *Freesound* dataset [11]. In our study, we selected all classes in the urban category since they are commonly involved in people's daily activities. The selected classes were *Helicopter, Chainsaw, Siren, Car horn, Engine, Train, Church bells, Airplane, Fireworks, Hand saw*, totaling 400 audio clips.

Degrading Method	δ = 50%	δ = 70%	δ = 90%
Dropping	-0.59	-0.07	-0.19
Cloning	+0.33	-0.11	-0.33
Replacing	-0.60	-0.77	+0.35

Table 4: Change of averaged 5-fold accuracy (in %) based on the ESC-50 [32] audio frames when comparing the results of audio degradation to the results without any degradation (50.16%). Similar to the field tests, the overall performance remains similar even when the audio is severely degraded.

Similarly, the audio data was split with a 5-fold basis. Before processing, the audio was re-sampled to 16KHz mono with 16-bit depth. We then extracted 19 MFCC features for every 50-ms window without overlaps. The audio was then degraded based on the same set of frames. Here we did not repeat the two paradigms as in the previous tests and we degraded both the training and test data. The training and test audio is from different audio clips, so there is no potential overlaps between the training and test sets.

The classifier is also a sequential neural network. Based on initial tuning on the raw audio, the network consists of two convolutional layers and four fully connected layers including the final output layer. There are 64 channels each for the convolutional layers and the filter size is 4 with a stride of 2. We added 0.1 dropout [35] and batch normalization [15] to both layers. The number of neurons is 64, 64 and 128 for the fully connected layers except for the output. The dropout of the fully connected layers is 0.3, and there is no batch normalization for them. All layers are activated by the ReLU function except for the output which is activated by the softmax. We used the cross entropy loss and the stochastic gradient descent optimizer with 0.01 learning rate and 0.9 momentum. The learning rate was dropped by a factor of 0.5 for every 3 steps whenever there was no improvement on validation. The network was developed using Python Keras [5].

5.4 Online test results

Table 4 shows the recognition performance of the ESC audio frames. The averaged 5-fold accuracy with no audio degradation is 50.16% with variance of 22.15%. The recognition performance obtained is comparable to some similar prior attempts [14, 32]. Similar to the field tests, recognition of the audio frames is also not affected by the implementation of the audio degradation. We also notice that dropping the frames does not harm the performance as in the field tests. That is possibly because the variability of the ESC audio is generally less than the field audio, and thus it requires less data to well-train a classifier.

6 RECOGNITION PERFORMANCE ON AUDIO SEGMENTS

Rather than studying just the frame-level features, we further examined the effects applying segment-level spectrogram for audio recognition. The intuition of this study is that frame degradation can potentially disrupt the temporal distribution of the original audio, so we hope to explore how the effects can be when integrating the degrading methods with segment-wise audio features.

6.1 Implementation

We still leveraged the same audio clips from the ESC-50 dataset [32] for the study. To extract segment-level spectrogram, each 5-second audio clip is windowed at the size of 1024 ms with 512 ms hops. The formulated spectrogram is in two dimensions with 128 mel bins along the time axis. We then developed our audio recognition framework based on a deep learning-based feature extractor proposed by Kumar et al. [18]. The feature extractor is a 19-layer convolutional neural network pre-trained on the Audio Set [12] weakly labeled data. The network takes as input 2D spectrogram and we extracted embedding features from the 16th layer of the network as the segment representations. Each embedding feature is in 512 dimensions. The features are then fed to a classifier.

In our tests, we applied both the random forest (RF) classifier and a neural network (NN). Both classifiers are fine-tuned based on 5-fold cross validation using the raw audio clips and are fixed afterwards. The RF consists of 600 estimators with gini classification criterion. It was developed with the Python Scikit-learn package [30] and all other parameters were left as default. The NN consists of three convolutional layers and four fully connected layers. Each convolutional layer is connected with 0.1 dropout [35] and batch normalization. The number of channels is all 32 and the filter size is 2 with single-step stride. The fully connected layers are of 128 neurons each with 0.3 dropout except for the output layer. All layers are activated by the ReLU activation. We used the categorical cross entropy loss and the stochastic gradient descent optimizer with 0.001 learning rate and 0.9 momentum. The learning rate was dropped by a factor of 0.5 for every 5 steps when there was no validation improvement. Further, the NN was developed with Keras [5].

To degrade the audio, we used a similar degrading window size of 50 ms with the same hop length. The neighboring range of replacement was 10 frames per side for the replacement-sampling method. The audio clips were then distorted before the computation of their spectrogram. Similar to the frame-level studies, we tested with only the valid audio mitigation levels (δ = 50%, 70%, 90%). We then reported the highest mean of the 5-fold accuracy in each test case.

6.2 Results

Table 5 presents the recognition performance with the segment spectrogram. The 5-fold accuracy values based on the original sounds are 79.50% and 81.25% for the RF and NN respectively. The corresponding 5-fold variance is 22.25% and 47.50%. The overall recognition performance is comparable to the results obtained by Kumar et al. [18].

First of all, we can see that the recognition accuracy drops in all test cases. It shows that the disruptions of the audio frame distributions can leave non-negligible impact on the recognition of the spectrogram. This is especially true when the audio frames are overly down-sampled. When δ = 70% in the down-sampling case, for example, the classification accuracy for the RF and the

Audio Processing Method	δ = 50%	δ = 70%	δ = 90%
Frame Dropping + RF	-5.75	-8.50	-22.50
Frame Dropping + NN	-9.25	-20.00	-48.50
Frame Cloning + RF	-7.10	-7.50	-5.75
Frame Cloning + NN	-5.50	-6.00	-5.00
Frame Replacing + RF	-5.25	-5.25	-6.00
Frame Replacing + NN	-6.75	-7.00	-8.25

Table 5: Change of averaged 5-fold accuracy (in %) based on the ESC-50 [32] segment spectrogram when comparing the results with audio degradation to results without any degradation (79.50% for RF and 81.25% for NN). Generally speaking, the classifiers are still able to classify the audio clips after the frames of each clip are re-sampled, but the global recognition performance drops from the original in all tested cases.

NN drops to 71.00% and 61.25% respectively. The reason can be that the spectrogram becomes less distinguishable when there are fewer frames within the segment. Due to the model complexity, the neural network tends to be even more sensitive to such frame loss. Secondly, both classifiers may still be able to recognize the audio segments despite the mitigation effects. As in the table, the accuracy values for δ = 70% in the replacement-sampling / up-sampling tests are 74.25% / 72.00% for the RF, and 74.25% / 75.25% for the NN. If our goal is to perform global recognition of the classes, we can see that the accuracy values based on the segment recognition are still promising when comparing to the results of the frame-level studies, even after degradation. This is due to the advantage of segment spectrogram and neural network embedding in interpretation of complex sounds. In other words, the degrading methods may still be applied to audio recognition when using segment-level features, yet we need to optimize the choice of the audio features and recognition methods for the best balance between privacy mitigation effects and recognition performance.

7 DISCUSSIONS AND FUTURE WORK

7.1 Discussions

From the Mechanical Turk studies, we can see that degradation of the audio at a strong level does help to mitigate people's privacy concerns towards the sensed audio in the given scenarios. From their responses, people are less concerned generally because the audio is mostly not intelligible when high proportion of the frames are re-sampled. However, some people are still worried about the risks potentially brought by the sensing processes, even with data degradation.

Combining results of the field tests and tests with the public online data, we see that the frame recognition accuracy remains generally unchanged from the base cases with no audio degradation to the tested cases. This is actually intuitive since the frame re-sampling processes do not disrupt the patterns within a frame, and hence the classifier can still distinguish the frames. The exception is when the audio is down-sampled at an extremely high level ($\delta \geq 90\%$) for the field data. This is possibly because the variability of the field data is higher than that of the online data, and the learning process is affected when there is not sufficient data for generalization. The extra tests when only the training sets are degraded further confirm our findings that learning and generalization of the classifier at a global level is not affected by individual frame re-sampling, and the audio recognition can be done across datasets with and without degradation. Hence, the audio frame degradation methods provide good opportunities for reliable sound frame recognition with effective mitigation of resulted privacy concerns.

From the results of the segment-level tests, we can see the trade-offs between sound privacy mitigation and the target class recognition. Potential factors such as the disruption of the frame distributions inside the spectrogram and change of the spectrogram resolution due to frame re-sampling may all hurt the classification results. Dropping the frames leads to the worst accuracy since it was implemented on individual segments and the valid segment size becomes much smaller when too many frames are removed. Correspondingly, the resolution of the spectrogram is lowered and it becomes more difficult to distinguish the patterns. However, we see that most of the segment-level accuracy is still over 70% on the ESC-50 data despite the performance drop from the raw test cases. Given the success of segment-level features over pure frame-level features such as the MFCCs for better interpretation of complex sounds, it can still be reasonable to leverage such features with audio degradation sometimes. In other words, the choices of the audio features, classification methods, and audio degradation parameters should be determined in a case-by-case manner in actual deployment.

7.2 Future work

In our study, the privacy mitigating effects are characterized based on an online study with pre-defined sensing scenarios and recorded sounds. However, privacy attitudes can be tied to several factors. For example, people could be more concerned if they realize that their own voice is being recorded even if the speech is unintelligible. People may also have different attitudes towards different sensing situations, especially those more sensitive ones in their daily living. Due to the scale of our study and the limitations of the online platform, we were not able to further explore such factors and they are left as future directions.

Both the audio frame degradation methods and our audio classification frameworks are prone to several limitations or assumptions. For example, the local device processing the audio frames needs to be trusted. Also, segment-level recognition of sounds has been studied for long and includes a large body of approaches. The actual trade-offs between privacy mitigation and audio recognition performance with different sensing and recognition methods should be further studied in the future.

8 CONCLUSIONS

In this work, we studied the extent to which frame-level audio degradation can mitigate people's privacy concerns with regards to acoustic sensing in mobile applications. Specifically, we designed and conducted an online survey with 4 common sensing scenarios by collecting responses from 266 participants using the Amazon

Mechanical Turk platform. We then analyzed the impact of audio degradation in audio-based human activity and context recognition with data collected in the home of 14 participants and from an online dataset. The effects on audio recognition performance were quantified by tests with frame-level and segment-level features and at different levels of degradation. Given the findings of our studies, degradation of audio frames shows a promising direction in privacy protection in a wider range of sound sensing and audio-driven human activity recognition domains.

REFERENCES

[1] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–28.

[2] Vincent Becker, Linus Fessler, and Gábor Sörös. 2019. GestEar: combining audio and motion sensing for gesture recognition on smartwatches. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 10–19.

[3] Francine Chen, John Adcock, and Shruti Krishnagiri. 2008. Audio privacy: reducing speech intelligibility while preserving environmental sounds. In *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 733–736.

[4] Jianfeng Chen, Alvin Harvey Kam, Jianmin Zhang, Ning Liu, and Louis Shue. 2005. Bathroom activity monitoring based on sound. In *International Conference on Pervasive Computing*. Springer, 47–61.

[5] François Chollet et al. 2015. Keras. https://keras.io.

[6] Delphine Christin, Andreas Reinhardt, Salil S Kanhere, and Matthias Hollick. 2011. A survey on privacy in mobile participatory sensing applications. *Journal of systems and software* 84, 11 (2011), 1928–1946.

[7] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one* 8, 3 (2013), e57410.

[8] Wenrui Diao, Xiangyu Liu, Zhe Zhou, and Kehuan Zhang. 2014. Your voice assistant is mine: How to abuse speakers to steal information and control your phone. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*. 63–74.

[9] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. 2018. Mitigating Bystander Privacy Concerns in Egocentric Activity Recognition with Deep Learning and Intentional Image Degradation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 132.

[10] Antti J Eronen, Vesa T Peltonen, Juha T Tuomi, Anssi P Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi. 2006. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 1 (2006), 321–329.

[11] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*. 411–412.

[12] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[13] Mark S Granovetter. 1977. The strength of weak ties. In *Social networks*. Elsevier, 347–367.

[14] Muhammad Huzaifah. 2017. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. *arXiv preprint arXiv:1706.07156* (2017).

[15] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[16] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. 2009. Exploring privacy concerns about personal sensing. In *International Conference on Pervasive Computing*. Springer, 176–183.

[17] Sacha Krstulović. 2018. Audio event recognition in the smart home. In *Computational Analysis of Sound Scenes and Events*. Springer, 335–371.

[18] Anurag Kumar, Maksim Khadkevich, and Christian Fügen. 2018. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 326–330.

[19] Sumeet Kumar, Le T Nguyen, Ming Zeng, Kate Liu, and Joy Zhang. 2015. Sound shredding: Privacy preserved audio sensing. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*. ACM, 135–140.

[20] Nicholas D Lane, Petko Georgiev, and Lorena Qendro. 2015. DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 283–294.

[21] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. In *The 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, 213–224.

[22] Eric C Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak N Patel. 2011. Accurate and privacy preserving cough sensing using a low-cost microphone. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 375–384.

[23] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 17.

[24] Daniyal Liaqat, Ebrahim Nemati, Mahbubur Rahman, and Jilong Kuang. 2017. A method for preserving privacy during audio recordings by filtering speech. In *2017 IEEE Life Sciences Conference (LSC)*. IEEE, 79–82.

[25] Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. 2009. SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*. ACM, 165–178.

[26] Md Nasir, Wei Xia, Bo Xiao, Brian Baucom, Shrikanth S Narayanan, and Panayiotis G Georgiou. 2015. Still together?: The role of acoustic features in predicting marital outcome. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[27] Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)* 33, 1 (2001), 31–88.

[28] Alexandru Nelus, Janek Ebbers, Reinhold Haeb-Umbach, and Rainer Martin. 2019. Privacy-Preserving Variational Information Feature Extraction for Domestic Activity Monitoring versus Speaker Identification. *Proc. Interspeech 2019* (2019), 3710–3714.

[29] M Pathak. 2010. Privacy Preserving Techniques for Speech Processing. *Dec* 1 (2010), 1–54.

[30] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[31] Pablo Perez Zarazaga, Sneha Das, Tom Bäckström, VV Raju, Anil Vuppala, et al. 2019. Sound Privacy: A Conversational Speech Corpus for Quantifying the Experience of Privacy. In *Interspeech*. International Speech Communication Association.

[32] Karol J. Piczak. [n.d.]. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia* (Brisbane, Australia, 2015-10-13). ACM Press, 1015–1018. https://doi.org/10.1145/2733373.2806390

[33] Andrew Raij, Animikh Ghosh, Santosh Kumar, and Mani Srivastava. 2011. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 11–20.

[34] Mirco Rossi, Sebastian Feese, Oliver Amft, Nils Braune, Sandro Martis, and Gerhard Tröster. 2013. AmbientSense: A real-time ambient sound recognition system for smartphones. In *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. IEEE, 230–235.

[35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[36] Edison Thomaz, Cheng Zhang, Irfan Essa, and Gregory D Abowd. 2015. Inferring meal eating activities in real world settings from ambient sounds: A feasibility study. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 427–431.

[37] Nik Vaessen. 2019. Jiwer. https://pypi.org/project/jiwer.

[38] Korosh Vatanparvar, Viswam Nathan, Ebrahim Nemati, Md Mahbubur Rahman, and Jilong Kuang. 2019. A Generative Model for Speech Segmentation and Obfuscation for Remote Health Monitoring. In *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 1–4.

[39] Danny Wyatt, Tanzeem Choudhury, and Jeff Bilmes. 2007. Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *Eighth Annual Conference of the International Speech Communication Association*.

[40] Koji Yatani and Khai N Truong. 2012. BodyScope: a wearable acoustic sensor for activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 341–350.

[41] Tomoko Yonezawa, Naoki Okamoto, Hirotake Yamazoe, Shinji Abe, Fumio Hattori, and Norihiro Hagita. 2011. Privacy protected life-context-aware alert by simplified sound spectrogram from microphone sensor. In *Proceedings of the 5th ACM International Workshop on Context-Awareness for Self-Managing Systems*. ACM, 4–9.