Extracting Speech from Motion-Sensitive Sensors

Chapter · December 2020

DOI: 10.1007/978-3-030-66172-4_9

Safaa Azzakhnini
6 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Ralf C. Staudemeyer
University of Applied Sciences Schmalkalden
34 PUBLICATIONS **414** CITATIONS
SEE PROFILE

Extracting speech from motion-sensitive sensors

Safaa Azzakhnini[1] and Ralf C. Staudemeyer[2]

[1] Mohammed V University, Faculty of Sciences, Morocco
[2] Schmalkalden University of Applied Sciences, Germany
safae.azzakhnini@gmail.com
r.staudemeyer@hs-sm.de

Abstract. The increasing presence of wireless sensor networks and the blanket re-use of the resulting data volumes by AI-based systems raises pressing ethical questions about the impact of these technologies on our society. One of the commonly used technologies is Smart Phones and similar mobile communication devices. These devices are equipped with rich sensors that provide an advanced and comprehensive user experience. However, it is a well-known problem that the presence of numerous sensors is of major concern to the privacy of users and their social environment. Previous studies already revealed that motion-sensitive sensors actually react to human speech. In this regards, Deep Neural Networks (DNN) proved very successful to model high-level abstractions in data. Our main focus is highlighting (i) the potential risks related to these sensors leaking private information about speech and (ii) the ethical implications of advances in (deep) machine learning as a threat to privacy. We showcase a simple attack in which collected data from accelerometer and Vibration Energy Harvester (VEH) sensors can be used to eavesdrop on speech. We propose a multistage stacked auto-encoder model that learns time and frequency features. We demonstrate the efficiency of our model with poor quality data and a very low sampling rate. We investigated three classification tasks: gender identification (i), hotwords detection (ii), and (iii) recognition of simple phrases selected. Our results confirm that motion-sensitive sensors are a rich source of personal data, from which highly sensitive information about people in close proximity to the sensor emerges.

Keywords: mobile devices · users privacy · deep neural networks· data fusion

1 Introduction

Privacy is increasingly a concern in today's digitally connected world. Personal data is being collected and stored in several daily used devices such as smartphones, mobile devices and wearables. These devices are often equipped with sensors to provide services based on the according sensor readings, like location, movement, temperature, and alike. The data from such sensors, however, can also be used for other purposes. Meanwhile, machine learning is a field of research that became a core component of many real-world applications in many domains including health, transportation, energy, education, banking, biometrics to cite a few. The diverse and large availability of data, coupled with rapid technological advances in machine learning algorithms (which learn from data), is changing society markedly. Although it enables the development of many tools with the potential of bringing good to society, their misuse might also generate or inflate risks that harm society and the private life of individuals [27,4,7]. In this work, we address the privacy concern raised when the combination of unprotected data collection and advanced learning algorithms may lead to non-transparent inferences. We mainly focus on data collected from sensors built-in mobile and wearable devices. We investigate how data created by a movement sensor and an energy harvesting component can be used to extract information about human voice communication.

Sensor data has already received remarkable attention from the security research community, as to better understand the potential impact of this data on user privacy. Projects have investigated opportunities to identify and track users [23,11,19,3]. This was investigated in particular with the acceleration sensor [32,28,25,31,38,8,14,34]. These studies did also show that the motion sensors included in smartphones are sufficiently sensitive to allow the identification of acoustic information based on the readings induced by sound waves. An according investigation of data provided by gyroscopes was performed by [24]. This paper was inspired by works discussing such effects of acoustics on gyroscope measurements [9,12,13]. The authors there demonstrated by a rich experimental study that gyroscope data is sufficiently sensitive to extract information about the original audio signal. This included the identification of the speaker's gender as well as an isolated hot-word. In [37] the authors investigate accelerometer data for hot-word detection. The main motivation is to enable accurate low-energy and low-cost implementation of voice control by using the accelerometer instead of the microphone. The obtained accuracies were competitive with voice control applications such as "Google Now" and "Samsung S Voice". However, mobile operating systems limit the sampling rate (usually to 200Hz). Low sampling rates pose a hard limit on the available data and therefore are a significant challenge to speech reconstruction. To overcome this challenge, a recent work [17] has proposed an eavesdropping attack by leveraging a distributed form of time-interleaved analog-digital-conversion to approximate a higher sampling rate. Combining the data provided by a geophone, an accelerometer, and a gyroscope they were able to reconstruct intelligible speech. A threat analysis of extracting speech signals from motion sensors of smartphones is provided by [1]. The authors there examined the presence of speech information in accelerometer and gyroscope data by studying many possible attack scenarios and analysing the behaviour of these sensors. Furthering this track of investigations a recent publication explored vibration energy harvesters (VEH) and whether they can be used like a sensor [20]. VEHs convert physical movement into electric energy, often to extend battery life. Because of the high availability of vibration sources, VEHs are considered an effective energy harvesting option for low-power mobile devices, like for the Internet of Things (IoT). The authors there also notice that VEHs can be sensitive enough to detect hot-words in speech. In other recent works, the authors propose an eavesdropping attack based on accelerometer readings. The authors in [2] studied an accelerometer-based speech recognition under the setup that the accelerometer is on the same smartphone as the speaker. Moreover, the authors in [5] design a deep learning-based framework to recognize and reconstruct speech signals only from accelerometer measurements.

Contributions In this paper, we propose to use DNNs on sensor data to increase the accuracy in recognising voice patterns. We want to determine the potential risks related to privacy when such data is not protected. Here we focus on the data collected from a VEH and an accelerometer while users were speaking as collected in a preceding study by [20]. Our intentions are twofold: Firstly we want to highlight the improved ability to extract acoustic patterns from sensors not primarily used for acoustic information by using DNNs. Secondly, we want to explore if the combination of data from different sensors can significantly improve the detection rates. We perform an experiment by using a DNN based on a stacked auto-encoder upon the collected data. Our model extracts acoustic features by exploring both time and frequency representations. The extracted features are then used in supervised classification to identify the speaker's gender, detect a simple hot-word, and distinguish it from short sample phrases. We show how the combination of the data provided by the VEH with the data of the acceleration sensor significantly improves the recognition rates in comparison to [20]. To that end we train the DNN with both sources. To the best of our knowledge the effect of speech on motion sensors has so far only been performed using manual feature extraction techniques [37,18,1,20,24,2] or using only one-way sensor [5]. We, therefore, assume our approach of using deep neural networks with different types of information in this context is novel.

Threat model When sensor readings expose voice communication additional threats to privacy become apparent. The threat model changes as an attacker then only needs access to sensor readings instead of the microphone directly. The attack vector thereby is extended to any application having access to the readings of relevant sensors, as for example on the user's device in figure 1.

The attacker here can identify acoustic patterns in the accelerometer and VEH readings. With the sensors used in the experiment, the user needs to be physically close to the device [20]. However, this seems very likely when using a mobile phone or a smart watch, which also happen to be the devices where accelerometer and VEHs are (to be) used. Furthermore, the attacker does not necessarily require direct access to the sensor data. We assume that access to locally cached sensor readings may be sufficient to allow offline attacks. We consider two scenarios in our study: In the first scenario, the attacker only has access to the data of one source. With the available data, we can compare having access to only the accelerometer or only the VEH. We determine how well the DNN identifies the speaker's gender, detect the hot-word, and identify short

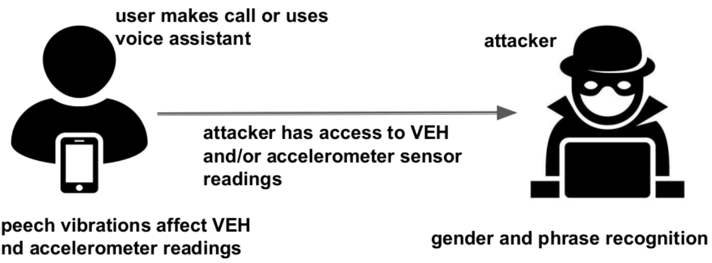
user makes call or uses voice assistant
attacker
attacker has access to VEH and/or accelerometer sensor readings
speech vibrations affect VEH and accelerometer readings
gender and phrase recognition

Fig. 1: example for an attack scenario

sample phrases for each of those data sets. In the second scenario, we assume the attacker to have access to both, the accelerometer and the VEH. Here we combine the data sets of the accelerometer and the VEH to train the DNN. We can compare the results with those derived by using a single data source.

2 Background

This section describes the design of the investigated sensors and provides background information about the typical architecture of an auto-encoder and the corresponding learning algorithm.

Vibration Energy Harvester (VEH) A VEH is a transducer that converts kinetic energy from vibrations to electrical power. For low-power electronic devices in specific environments, they can harvest enough energy to operate the device [10,15,30]. A VEH can be seen as having three parts: the transducer to convert the kinetic to electrical energy, a power-electronic interface, and some electrical energy storage, like a battery [29]. Common VEH transducers are piezoelectric, as this type has shown the highest potential for harvesting energy [22,36]. Suitable vibration sources are diverse, as for example human motion, waves, wind, or vibrations of machinery. A typical piezoelectric element as used in VEHs is illustrated in figure 2 where one end of a cantilever beam is fixed to the device, while the other is set free to oscillate (vibrate). When the piezoelectric is affected by vibrations, an AC voltage is generated by the accumulation of positive and negative charges on the two opposing sides. The AC voltage generated in general is proportional to the applied stress.

Acoustic effect Sound waves, when emitted, are moving through air and cause pressure on the cantilever beam. Experiments [20] have demonstrated this effect by having a person shout three times while physically near to the piezoelectric part. The generated signal (see figure 3) shows how the VEH's voltage peaked with each shout.

Accelerometer Accelerometers turn acceleration into an electrical signal based on the same operating principles as VEHs. The acceleration in different dimensions can be translated to changing positions. Raw gyroscope data consist of

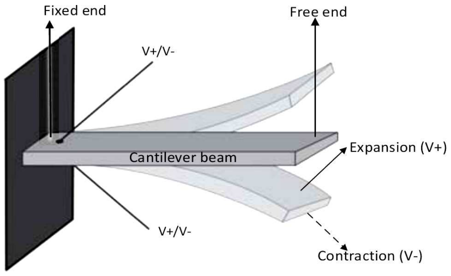

Fig. 2: piezoelectric transducer [20]

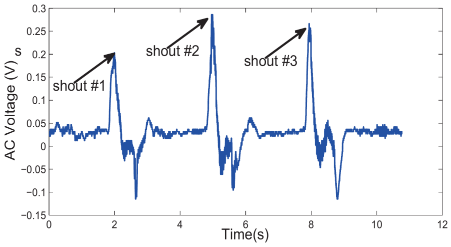

Fig. 3: effect of shouting on VEH piezoelectric cantilever beam [20]

three values indicating the acceleration along the x-axis, y-axis, and z-axis (usually corresponding to the up-down, right-left, and front-back movement respectively).

Acoustic effect Recent work [37] showed that accelerometers are sensitive enough to draw conclusions about human speech. The authors there recorded sensor output while a speaker was spelling the vowel “A”. The spectrum analysis of the output signal shows a considerable variation of the accelerometer readings during speech. They reported that the human voice has sufficient sound pressure to have detectable impact on smartphone accelerometers.

Stacked Auto-Encoders An auto-encoder (AE) is a type of artificial neural network (ANN) for unsupervised learning. The learning objective of the AE is to map the data of the input layer to the output layer in the way it is desired. The result is an approximation of the so-called identity function, where the output is a representation of the input. The architecture of an AE divides the ANN into an encoder and a decoder. The encoder takes the data at the input neurons and creates a “restricted” representation of it at the hidden layer. Since the hidden layer is smaller than the input layer it learns only the most relevant aspects of the input. The decoder then tries to reconstruct the original input from the representation in the hidden layer. This produces a higher-level representation from the lower-level representation of the input [6].

A stacked auto-encoder (SAE) is an ANN consisting of multiple hidden layers creating a deep neural network architecture. The SAE applies the so-called greedy layer-wise pre-training strategy which addresses the error-causing vanishing gradient problem. In SAEs the input layer is the encoded layer trained on the raw input. The output then is used as input to the next AE to obtain the next encoded layer and this process is repeated for subsequent layers. Stacking layers like this can then lead to deep stacked auto-encoders that carry some of the interesting properties of deep models [6].

3 Learning Acoustic Information

Here we present an overview of our proposed approach and discuss details, before presenting the experiments in the next chapter.

Classification task Classification tasks are tied into information representation. The learning process on how to represent data is a critical step that on the one hand should preserve as much information as possible from the input data. On the other hand this process should eliminate redundancies to foster the extraction of structures and properties. Motion sensors are designed to respond to movement. Their output signals originate from physical movement of an accordingly designed part of the sensor. They are particularly used for tasks related to motion recognition, such as identifying physical activity. Modelling motion sensor data to perform sound recognition is an especially challenging task, because the sensor's sensitivity is optimised towards such movements and not sound.

The artificial learner requires preprocessed data in form of features to learn from. A feature is a measurable property fed to the learning algorithm. These are normally manually extracted relying on knowledge of a human expert. This expertise is domain- and/or sensor-specific and is required for each new dataset or sensor modality in order to engineer the suitable features for a specific application. Therefore, the use of manual feature extraction is very limited. Furthermore it cannot be generalised across different application domains. As accelerometer and VEH sensors are of a non-acoustic nature, we do not benefit from a prior knowledge about useful measurements to apply in order to extract the acoustic information. To deal with this issue, we propose an unsupervised deep learning approach to automatically learn suitable features without relying on handcrafted features. Here features are automatically extracted from data through layers were each successive layer acts as a feature extractor and is hypothesized to represent the data in a more abstract way. This process is unsupervised, which means that it is independent to a specific classification task. To this end we propose a SAE to discover relevant complex structures underlying speech and to learn a deep and high-level representation robust to intra-class variability including the sensor direction and the speaker speaking. An architectural overview of our approach is illustrated in figure 4. It is divided into two main phases: unsupervised feature learning, where we added the combination of the data-sources as well, and supervised classification.

Feature learning In this step, we investigate time and frequency data separately to train a bimodal representation from each sensor. We perform Fourier transformations on the frequency data. Then we use the greedy layer-wise training for the SAE. Therein, the features learned in a hidden layer are used as input to the next AE in order to produce a new representation of the data. By representing the data through layers we enable learning of complex patterns across data variations. After extracting the features separately from each source, that is time and frequency, we combine them into a joined time-frequency representation. This joint representation leads to a shallow model, thereby making it difficult for a single hidden layer model to directly find correlations between representations that have been joined. We, therefore, again apply greedy layer-wise

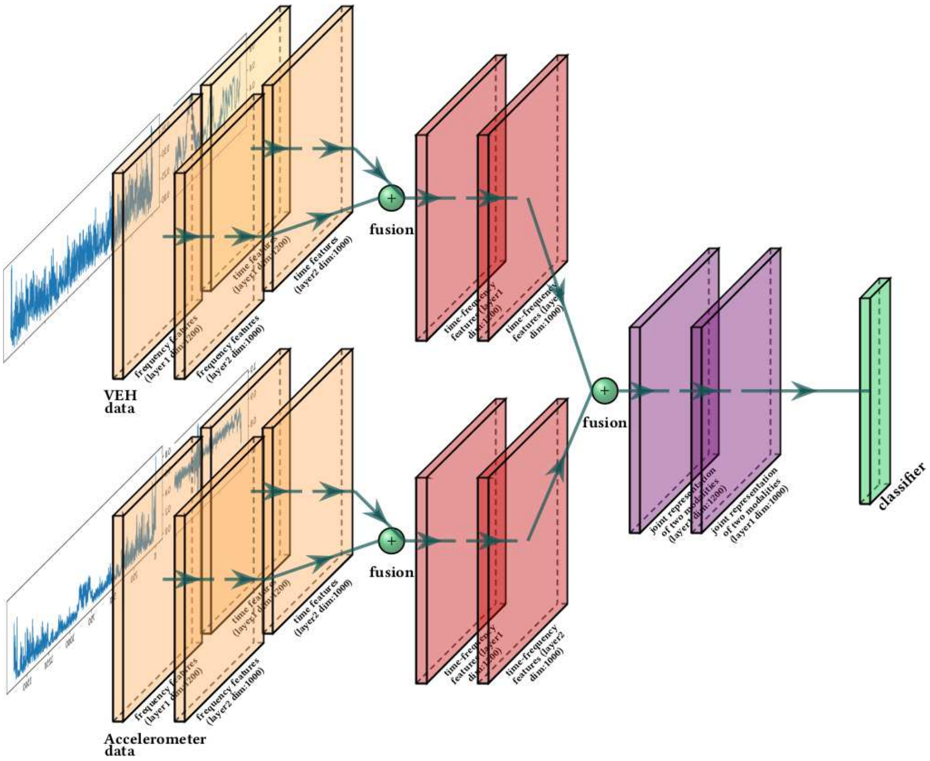
frequency features (layer1 dim:1200)
frequency features (layer2 dim:1000)
time features (layer1 dim:1200)
time features (layer2 dim:1000)
VEH data
fusion
time-frequency features (layer1 dim:1200)
time-frequency features (layer2 dim:1000)
frequency features (layer1 dim:1200)
frequency features (layer2 dim:1000)
time features (layer1 dim:1200)
time features (layer2 dim:1000)
Accelerometer data
fusion
time-frequency features (layer1 dim:1200)
time-frequency features (layer2 dim:1000)
fusion
joint representation of two modalities (layer1 dim:1200)
joint representation of two modalities (layer1 dim:1000)
classifier

Fig. 4: The figure shows the whole network architecture. The two inputs are the frequency and time representation of the VEH and the accelerometer data. Each layer correspond to the hidden layer encoded using the autoencoder. The layers are stacked using layer wise training strategy. The last layer represents the classifcation step performed after the unsupervised features learning from previous layers.

training to improve discovery of high-level correlations across the two representations.

Data From Multiple Sources With the assumed availability of different sensors, we then have separate types of data sources about a given moment. The machine learning community assumes potential in improved learning algorithms to specifically exploit such multi-modal data to form a unified picture [26]. Modelling speech recognition from data of non-acoustic sensors is challenging. Additional problems to tackle are the limited sampling frequency and the interference from the device's original function (detecting movement, harvesting energy) with our intended function (detecting spoken language). Our study specifically aims to determine to what extent combining data provided by different sensors can provide improved results. To that end, our multi-layer approach combines separately trained models into a joint representation.

Supervised classification We then use supervised classification on the extracted features. For this the fused representation functions as the input, thus providing features across the original data sources. We repeat the three classification tasks for the speaker's gender identification, the hot-word detection, and the recognition of the sample phrases.

4 Experimental study

In this section we are describing our experiments in more detail. We start by presenting the available data and then explain how we pre-processed it and performed the feature learning and classification.

Data The dataset we used is described in more detail in [20]. It is the only work we are aware of that already studied the potential of detecting acoustic information from VEH data. It contains the data for both, a VEH and an accelerometer, while different persons performed identical tasks repeatedly. Involved were eight individuals, four being male and four female, and the experiments were performed with two different orientations of the devices (horizontal and vertical). The devices were positioned close to the persons (3 cm) and the experiments repeated 30 times for the hot-word "Ok Google" and at least ten times for the phrases "Good morning", "how are you", and "fine thank you". Overall the data-set contains 1155 samples. Figure 5, represents the accelerometer and VEH sensor outputs while a person spoke the four phrases.

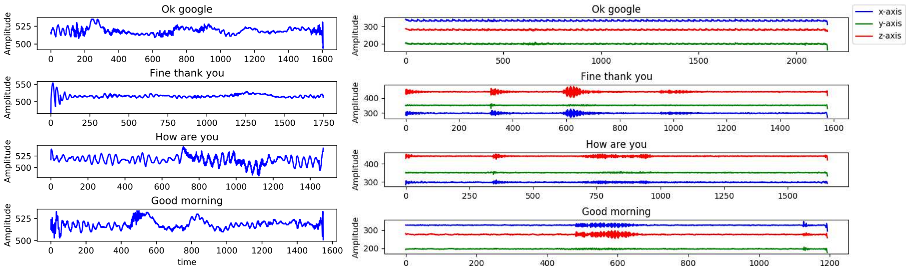

Fig. 5: The VEH signal (left) and the accelerometer outputs (x axis,y-axis, z-axis on right) while the user is speaking the four phrases (”good morning”, ”okay google”, ”fine thank you” and ”how are you”)

Preprocessing In the pre-processing we apply our domain knowledge to address the specifics of the different data sources. As we face varying lengths of the samples, we started by interpolating all the samples in the data to the mean length. We separately handle the temporal and frequency representations. To minimise the signal-to-noise ratio we filter the input signals and normalise them.

On the frequency representation we apply a fourier transformation and calculate the magnitude of the obtained complex values, which we then also normalised. For the accelerometer data, we down-sample the signal to 200Hz. This is the limit on sampling frequency as posed by the mobile operating systems Android and iOS. We then interpolate the samples to their mean length and normalize the data. After first learning features for each axis (x, y, and z), we then compute the magnitude over the dimensions to obtain an overall feature vector. The three acceleration channels were combined as one using square summing to obtain the magnitude acceleration, which is orientation independent.

Feature Learning and Classification The training procedure for time and frequency representations each is executed for 50 epochs, using a mini-batch size of 30 and learning rate of 0.001. The RMSprop variant of the stochastic gradient descent is used as the optimization algorithm. For the feature learning step, the used number of hidden units for the first layer is 1200 and for the second layer is 1000. For the classification we used and compared three common learning algorithms in order to select the good learner that can find out the relevant patterns from the obtained features from the Autoencoder. Principaly, Support Vector Machine (SVM) using the RBF kernel with $\gamma = 0.01$ and $C = 100$, K-nearest neighbours (KNN) with $K = 3$ and Neural Network Classifier (NN) with 2 hidden layers containing 100 units. The optimal hyperparameters were optimized by cross-validated grid-search over a parameter grid.

Evaluation In the evaluation we used a k-fold cross-validation with $k = 10$. For this we divided the data into k equal folds (portions). We then trained the model on the $k - 1$ folds and test it against the remaining folds. That process was repeated $k = 10$ times. The final performance after that corresponds to the average of the obtained values. We used cross-validation analysis to ensure that all data was used for both, training and test. The classification of the proposed framework was performed using the four metrics accuracy, precision, recall, and F-measure.

5 Results and discussion

In this section we present the results from our experiments in detail.

Single Data-Sets We first evaluate the results from using the data of the accelerometer or VEH on their own, each. Tables 1, 2 present the classification performance for each of our metrics, that is accuracy (acc), precision (prec), recall (rec) and F-measure (f-score), as determined for each of the classifications (gender identification, hot-word detection, and phrase recognition) for each of the used learning algorithms KNN, SVM, and NN (alg) for each of the data representations time-only, frequency-only, and our model. These allow us to see how our model compares to using only the time- or only the frequency-data. On the hot-word classification the KNN and SVM algorithms with our model both

Table 1: The obtained results (%) using only VEH data

		Hot-word detection				**Gender identification**				**Sentences recognition**			
alg.	**rep**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**
KNN	time	74	76	74	74	71	72	71	70	62	62	62	60
	freq	69	69	69	69	76	77	76	76	55	55	55	53
	our model	75	76	75	75	79	80	79	79	64	65	64	62
SVM	time	71	72	71	71	70	71	70	69	63	64	63	61
	freq	69	70	69	69	76	77	76	76	54	54	54	53
	our model	75	76	75	75	80	80	80	80	64	65	64	64
NN	time	70	72	70	70	62	64	62	61	61	62	61	60
	freq	65	68	65	63	72	74	72	72	54	54	54	51
	our model	75	76	75	75	77	79	77	77	65	65	65	64

Table 2: The obtained results (%) using only accelerometer data

		Hot-word detection				**Gender identification**				**Sentences recognition**			
alg.	**rep**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**
KNN	time	71	72	71	71	83	83	83	83	58	59	58	56
	freq	55	55	55	55	64	65	64	63	42	38	42	37
	our model	77	77	77	77	83	83	83	83	65	65	65	63
SVM	time	58	60	58	54	80	80	80	80	48	31	48	33
	freq	58	58	58	58	71	72	71	71	41	39	41	39
	our model	76	76	76	76	86	86	86	86	65	65	65	64
NN	time	53	55	53	48	61	65	61	58	47	23	47	30
	freq	54	56	54	50	66	68	66	65	45	35	45	34
	our model	68	70	68	67	80	81	80	79	54	57	54	50

achieved an accuracy of 75% when used on the VEH data and 76% when used on the accelerometer data, in all cases out-performing the use of only time- or frequency representations. For gender identification results, the best classification performance was achieved by our model in combination with SVM, having an accuracy of 86% using the accelerometer data and near 80% using the VEH data, again out-performing the use of only one data-representation. The accuracy of our model in recognizing the phrases was in the range of 64-65% for all combinations but using NN on the accelerometer data and once more out-performed the use of single data representations in all combinations. The F-score shows comparable values. Therefore, we can conclude that features learned from the joint representation of time and frequency information leads to a considerable improvement of the classifications. Both, frequency and time representations, contain important information that can be combined for better results here.

We assume that more abstract features have been learned in the process. We highlight that sufficient information about the original activity of shouting can already be extracted with relevant accuracy even when using only one of the data sources.

Table 3: The obtained results (%) combining the VEH and the Accelerometer data

		Hot-word detection				**Gender identification**				**Sentences recognition**			
alg.	**rep**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**	**acc**	**prec**	**rec**	**f-score**
KNN	time	78	79	78	78	81	81	81	81	69	69	69	67
	freq	69	70	69	69	79	80	79	79	55	55	55	52
	our model	83	83	83	83	88	88	88	88	73	74	73	72
SVM	time	76	76	76	76	79	79	79	79	66	66	66	64
	freq	73	73	73	73	81	82	81	81	59	59	59	58
	our model	86	86	86	86	91	92	91	91	77	78	77	77
NN	time	68	69	68	67	73	74	73	73	61	64	61	59
	freq	71	72	71	71	79	80	79	79	59	59	59	57
	our model	81	82	81	80	88	89	88	88	74	75	74	74

Combination of Data-Sets Next we examine if access to multiple data sources further increases the classifications. For this we repeated the training using both, the VEH and the accelerometer data, as described above. The results are — formatted as the previous tables — shown in table 3. Combining the data-sources has significantly increased the accuracy of the classifications across the board by around 10%. The highest F-scores of 91%, 85%, and 77% for gender identification, hot-words detection and recognition of phrases respectively, were achieved when using the SVM classifier with our model. The increase was higher for our model than if using only the time or the frequency representation. In each of its levels, the ANN must have learned additional correlations between the data variables across frequency and time representations. Overall, the joint representation has lead to remarkably improved accuracy.

In table 4 we compare our model with the results from the original work from [20] which only used the VEH data. The authors there compared results for different positions of the VEH. Recognising the importance of positioning, we specifically wanted the DNN to cope with this, as we hardly can influence the positioning in the scenario of spying. This way our results should be better suited to assess the practicality of an according attack vector. Moreover, we applied our model to the gyroscope data used by [24] for isolated words recognition. We compare our results with those obtained by the authors for the user independent case. The results show that our model provided better accuracy than the state of the art works, that were based on manual features.

method	**accuracy**
Hot-words detection using VEH in a horizontal position [20]	73%
Hot-words detection using VEH in a vertical position [20]	63%
Hot-words detection using VEH invariant to sensor orientation (our model)	75%
Hot-words detection combining VEH with accelerometer (our model)	**86%**
Isolated words recognition (11 words) using gyroscope with SVM (Speaker-independent) [24]	10%
Isolated words recognition (11 words) using gyroscope (our model) Speaker-independent case	**25%**

Table 4: comparison with state of the art methods

The results show that our deep auto-encoder approach can improve the recognition of acoustic patterns from non-acoustic sensors, here acceleration and voltage readings. We do not claim that the proposed approach represents a direct substantial risk to privacy, yet, as the data-set is small and was derived in a very specific setting. However, mobile sensors beyond the obvious microphone and camera could become targeted by attackers as they — as of today — are often less protected. Despite the fact that the data provided by such sensors is always cluttered due to their main purpose, it still is possible to draw conclusions on audio information from them by using two main approaches. The first is to combine data from multiple sources by considering a multimodal architecture. This will exploit the complementary between multiple modalities (information obtained from multiple sensors) and will lead to more accurate results. The second is to include a de-noising component in the autoencoder. In fact, de-noising data is one of the areas where auto-encoders have been most successful [35].

Discussion Considering that such sensors might generally not be considered as sensitive and therefore be less protected, they could rise to become popular attack surfaces in the future. Based on our results an attacker with access to the readings of multiple sensors must be regarded as dangerous, even if the primary function of the sensors seem harmless at first. With today's mobile devices many people already carry a multitude of sensors around and the trend seems to be for *even more.*

The findings compiled show that motion data are a rich source of personal data. The misuse of such data using learning algorithms that can extract visible and invisible patterns and correlations from it can lead to leakage of sensitive information such as the individual's speech. Furthermore, by combining different data sources we can learn more than from one source independently. Thus, more accurate information can be inferred from a combined analysis (in the studied case the accuracy has increased by 10%), which increases the risk of the privacy breach. Recognizing the speech does not only give information about what a speaker says, but also its attitude toward the listener and the topic under discussion, and the speaker own current state of mind as well. Many works have discussed the inferences that can be drawn from human speech extracted from audio data [21]. Therefore, the inferred speech information from motion sensors, in turn, can be used to deduce more non-transparent insights about individuals and manipulation about their private life [7]. Several examples of data breaches were revealed where personal information was exploited for many purposes, including political purposes [16], and others [33]. Therefore, the misuse of such data can seriously affect an individual's relationships, employability, or financial status, or lead to negative consequences for essential rights and social values such as freedom of expression, respect for private life. The privacy threat of unexpected inferences from unprotected data sources is not limited to those discussed in this paper. The problem of undesired inferences goes far beyond motion sensors and the deduced insights are related to the samples present in the used dataset. Thus, a larger database will contain a variety of characteristics from personal attributes in addition to gender and identity or age. Such attributes may include, emotions, personality traits, sexual orientation, ethnicity, religious and political views to cite a few. This diversity of data will allow discovering other correlations and obtaining more analysis. In sum, the aim of the paper is mainly to raise awareness about the ethical and privacy implications of the advancement of learning algorithms coupled with the growing availability of data. This is achieved by demonstrating how machine learning can be used as a tool for privacy breach and manipulation. Advances in technology change how personal information is collected and analysed, and therefore create new privacy risks. Thus the continued debate is needed to guide the development not only of technology but also of the policies that enable its use. And governments need to be more serious about finding a solution to limit the power that larger companies have over citizens.

6 Conclusions and future work

In this paper, we investigate the technical feasibility of speech inference from motion data using advanced machine learning models. we explore how non-acoustic sensor readouts can be used in uni-/multi-modal attacks. We propose a multi-level time-frequency based deep neural network to extract acoustic patterns from an accelerometer sensor and an energy harvesting component. Our model detects gender, single hot-words and spoken phrases with an accuracy of up to 91%, 85%, and 77% respectively. This findings show that motion sensors data are a rich source of personal data. They can be sufficient to obtain information about a device holder's speech especially when used data is combined from multiple sensors. By combining data sources we can learn more than would be the case from analysing single source independently, and more accurate predictions may be inferred which increases issues about privacy or decreases the privacy guarantee. An attacker with access to an accelerometer and some sensitive energy harvesting module is able to eavesdrop on human speech and draw conclusions about its content. Therefore, they could be considered private data in the same sense as audio data. The privacy of mobile device and wearables users, is a concern of growing importance. The zero permission nature of embedded motion sensors make acquiring the data easier. The collection of such data combined with a misuse of machine learning algorithms (which learn from data) can lead to a serious privacy risks and leakage of sensitive inferences about the user including his speech. The problem of undesired inferences goes far beyond motion sensors data and needs to be addressed for other data sources as well. Therefore, further research is required into the privacy implications of unprotected data collection taking into account the evolving state of the art in machine learning algorithms. Furthermore, a continued debate is needed not only about control over all sensor data, but also to guide the development of technology and of the policies that enable its use.

We consider the contents of this article to be early work on this topic. An interesting next step would be to examine the attack vector under real-world conditions. For further experiments a larger annotated data-set including more sensors as found in smartphones and a possibly large set of recorded situations would be needed. Only then would it be possible to realistically judge the threat that for example is posed by smartphones today, when installed applications are allowed access to sensors without care. Many factors usually do affect sensors and different sensors each have their specifics in how they are affected. This provides a large variety of possible experiments from recording and annotating data to performing analyses on that data then. Concerning the neural network it might be beneficial to use recurrent neural networks with long-short-term-memory to capture the temporal relationships in the data.

References

1. Anand, S.A., Saxena, N.: Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In: 2018 IEEE Symposium on Security and Privacy (SP). Vol. 00. pp. 116–133 (2018)
2. Anand, S.A., Wang, C., Liu, J., Saxena, N., Chen, Y.: Spearphone: A speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers. arXiv preprint arXiv:1907.05972 (2019)
3. Aviv, A.J., Sapp, B., Blaze, M., Smith, J.M.: Practicality of accelerometer side channels on smartphones. In: Proceedings of the 28th Annual Computer Security Applications Conference. pp. 41–50. ACM (2012)
4. Azencott, C.A.: Machine learning and genomics: precision medicine versus patient privacy. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **376**(2128), 20170350 (2018)
5. Ba, Z., Zheng, T., Zhang, X., Qin, Z., Li, B., Liu, X., Ren, K.: Learning-based practical smartphone eavesdropping with built-in accelerometer. In: Proceedings of the Network and Distributed Systems Security (NDSS) Symposium. pp. 23–26
6. Baldi, P.: Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML workshop on unsupervised and transfer learning. pp. 37–49 (2012)

7. Berendt, B.: Privacy beyond confidentiality, data science beyond spying: From movement data and data privacy towards a wider fundamental rights discourse. In: Annual Privacy Forum. pp. 59–71. Springer (2019)
8. Cai, L., Chen, H.: Touchlogger: Inferring keystrokes on touch screen from smartphone motion. HotSec **11**, 9–9 (2011)
9. Castro, S., Dean, R., Roth, G., Flowers, G.T., Grantham, B.: Influence of acoustic noise on the dynamic performance of mems gyroscopes. In: ASME 2007 International Mechanical Engineering Congress and Exposition. pp. 1825–1831. American Society of Mechanical Engineers (2007)
10. Choi, S., Seong, M., Kim, K.: Vibration control of an electrorheological fluid-based suspension system with an energy regenerative mechanism. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering **223**(4), 459–469 (2009)
11. Das, A., Borisov, N., Caesar, M.: Tracking mobile web users through motion sensors: Attacks and defenses. In: NDSS (2016)
12. Dean, R.N., Flowers, G.T., Hodel, A.S., Roth, G., Castro, S., Zhou, R., Moreira, A., Ahmed, A., Rifki, R., Grantham, B.E., et al.: On the degradation of mems gyroscope performance in the presence of high power acoustic noise. In: 2007 IEEE International Symposium on Industrial Electronics. pp. 1435–1440. IEEE (2007)
13. Dean, R.N., Castro, S.T., Flowers, G.T., Roth, G., Ahmed, A., Hodel, A.S., Grantham, B.E., Bittle, D.A., Brunsch, J.P.: A characterization of the performance of a mems gyroscope in acoustically harsh environments. IEEE Transactions on Industrial Electronics **58**(7), 2591–2596 (2010)
14. Dey, S., Roy, N., Xu, W., Choudhury, R.R., Nelakuditi, S.: Accelprint: Imperfections of accelerometers make smartphones trackable. In: NDSS (2014)
15. Donelan, J.M., Li, Q., Naing, V., Hoffer, J., Weber, D., Kuo, A.D.: Biomechanical energy harvesting: generating electricity during walking with minimal user effort. Science **319**(5864), 807–810 (2008)
16. Economist, T.: The cambridge analytica scandal – britain moves to rein in data-analytics (2018), `https://www.economist.com/britain/2018/03/28/britain-moves-to-rein-in-data-analytics`
17. Han, J., Chung, A.J., Tague, P.: Pitchin: Eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion (2017)
18. Han, J., Chung, A.J., Tague, P.: Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion. In: Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks. pp. 181–192. ACM (2017)
19. Han, J., Owusu, E., Nguyen, L.T., Perrig, A., Zhang, J.: Accomplice: Location inference using accelerometers on smartphones. In: 2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012). pp. 1–9. IEEE (2012)
20. Khalifa, S., Hassan, M., Seneviratne, A.: Feasibility and accuracy of hotword detection using vibration energy harvester. In: World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016 IEEE 17th International Symposium on A. pp. 1–9. IEEE (2016)
21. Kröger, J.L., Lutz, O.H.M., Raschke, P.: Privacy implications of voice and speech analysis–information disclosure by inference. In: IFIP International Summer School on Privacy and Identity Management. pp. 242–258. Springer (2019)
22. Lan, G., Xu, W., Khalifa, S., Hassan, M., Hu, W.: Veh-com: Demodulating vibration energy harvesting for short range communication. In: Pervasive Computing

and Communications (PerCom), 2017 IEEE International Conference on. pp. 170–179. IEEE (2017)

23. Matyunin, N., Szefer, J., Katzenbeisser, S.: Zero-permission acoustic cross-device tracking. In: 2018 IEEE International Symposium on Hardware Oriented Security and Trust (HOST). pp. 25–32. IEEE (2018)
24. Michalevsky, Y., Boneh, D., Nakibly, G.: Gyrophone: Recognizing speech from gyroscope signals. In: USENIX Security. pp. 1053–1067 (2014)
25. Miluzzo, E., Varshavsky, A., Balakrishnan, S., Choudhury, R.R.: Tapprints: your finger taps have fingerprints. In: Proceedings of the 10th international conference on Mobile systems, applications, and services. pp. 323–336. ACm (2012)
26. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 689–696 (2011)
27. Nissim, K., Wood, A.: Is privacy privacy? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **376**(2128), 20170358 (2018)
28. Owusu, E., Han, J., Das, S., Perrig, A., Zhang, J.: Accessory: password inference using accelerometers on smartphones. In: Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications. p. 9. ACM (2012)
29. Rao, Y., Cheng, S., Arnold, D.P.: An energy harvesting system for passively generating power from human activities. Journal of Micromechanics and Microengineering **23**(11), 114012 (2013)
30. Rome, L.C., Flynn, L., Goldman, E.M., Yoo, T.D.: Generating electricity while walking with loads. Science **309**(5741), 1725–1728 (2005)
31. Simon, L., Anderson, R.: Pin skimmer: Inferring pins through the camera and microphone. In: Proceedings of the Third ACM workshop on Security and privacy in smartphones & mobile devices. pp. 67–78. ACM (2013)
32. Song, C., Lin, F., Ba, Z., Ren, K., Zhou, C., Xu, W.: My smartphone knows what you print: Exploring smartphone-based side-channel attacks against 3d printers. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 895–907. ACM (2016)
33. Swinhoe, D.: The 14 biggest data breaches of the 21st century (2020), `https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html`
34. Van Goethem, T., Scheepers, W., Preuveneers, D., Joosen, W.: Accelerometer-based device fingerprinting for multi-factor mobile authentication. In: International Symposium on Engineering Secure Software and Systems. pp. 106–121. Springer (2016)
35. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research **11**(Dec), 3371–3408 (2010)
36. Vullers, R., van Schaijk, R., Doms, I., Van Hoof, C., Mertens, R.: Micropower energy harvesting. Solid-State Electronics **53**(7), 684–693 (2009)
37. Zhang, L., Pathak, P.H., Wu, M., Zhao, Y., Mohapatra, P.: Accelword: Energy efficient hotword detection through accelerometer. In: Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services. pp. 301–315. ACM (2015)
38. Zhang, Y., Xia, P., Luo, J., Ling, Z., Liu, B., Fu, X.: Fingerprint attack against touch-enabled devices. In: Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices. pp. 57–68. ACM (2012)