Feasibility and Accuracy of Hotword Detection using Vibration Energy Harvester

Sara Khalifa*‡, Mahbub Hassan*‡, Aruna Seneviratne†‡
*School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia
Email: {sarak, mahbub}@cse.unsw.edu.au
†School of Electrical and Telecommunication Engineering, University of New South Wales, Sydney, NSW 2052, Australia
Email: a.seneviratne@unsw.edu.au
‡National ICT Australia, Locked Bag 9013, Alexandria, NSW 1435, Australia
Email: {sara.khalifa, mahbub.hassan, aruna.seneviratne}@nicta.com.au

***Abstract*—Vibration energy harvesting (VEH) is a promising source of renewable energy that can be used to extend battery life of next generation mobile devices. In this paper, we study the feasibility and accuracy of VEH for detecting hotwords, such as "OK Google", used by popular voice control applications to distinguish user commands from other conversations. The idea of using power signals of VEH to detect hotwords is based on the fact that human voice creates vibrations in the air, which could be potentially picked up by the VEH hardware inside a mobile device. Using off-the-shelf VEH product, we conduct a comprehensive experimental study involving 8 subjects. We analyse two possible usage scenarios for the VEH hardware. In the first scenario, the user is not required to talk directly to the device (indirect), but the VEH is expected to pick up the ambient vibrations caused by user-generated sound waves. In the second, the user is expected to direct his voice to the VEH (direct) and talk to it from a close distance. For both usage scenarios, we evaluate two types of hotword detection, speaker-independent and speaker-dependent. We find that VEH can detect hotwords more accurately in the direct scenario compared to the indirect. For the direct scenario, our results show that a simple Decision Tree classifier can detect hotwords from VEH signals with accuracies of 73% and 85%, respectively, for speaker-independent and speaker-dependent detections. Finally, we show that these accuracies are comparable to what could be achieved with an accelerometer sampled at 200 Hz.**

***Index Terms*—Hotword detection, Vibration energy harvesting, Accelerometer, Voice control applications.**

I. Introduction

With increasing user demand for more power and functionality, manufacturers of mobile devices are forced to find new energy solutions beyond batteries. For it, there is a recent focus on vibration energy harvesting (VEH) as a viable option for mobile devices to generate electrical energy from ambient sources [1], [2]. VEH is considered one of the most effective energy harvesting options for the future internet of things due to the ubiquitous presence of vibration sources in the environment. Significant recent research confirms that VEH can harvest usable electric power for personal mobile devices by harnessing vibrations due to human motion [3]–[5]. These developments point to future mobile devices that will be equipped with some sort of VEH hardware to ease the dependence on batteries.

978-1-5090-2185-7/16/$31.00 ©2016 IEEE

Although the primary purpose of VEH is to generate electric power, in principle, it could also be used as a potential sensor to detect or identify the source of the vibration. The ability to detect the vibration source can lead to many potential applications for the VEH hardware beyond its primary use of energy harvesting. Indeed, recent work has convincingly demonstrated that VEH can be used as an effective sensor for human activity recognition due to the fact that different activities create different patterns of ambient vibrations, which produce different energy generation patterns in the VEH circuit [6].

In this paper, we study VEH's feasibility and accuracy for detecting hotwords, such as "OK Google", which are used by voice control applications to delineate user commands from background conversations. Because VEH does not require any power supply to operate, it offers unique power saving opportunity if used as a sensor for hotword detection. In contrast, embedded microphones used in mobile devices for hotword detection consume significant power [7]. To reduce power consumption of hotword detection, researchers have recently considered other low-power sensors, such as gyroscopes [8] and accelerometers [7], that consume less power than MEMS microphones. VEH's feasibility for hotword detection will open the door for further power saving opportunities in wearable devices.

The contribution of this paper can be summarized as follows:

- We conduct the first study to assess the feasibility and accuracy of VEH-based hotword detection.
- Using off-the-shelf VEH product, we conduct a comprehensive experimental study involving 8 subjects. Our experiments involve the analysis of two possible usage scenarios, indirect and direct. In the first, the VEH is only expected to pick up the *ambient* vibrations caused by user speech in the vicinity of the device. In the second, the user talks directly to the surface of the piezoelectric beam. For both usage scenarios, we evaluate two types of hotword detection, speaker-independent, which does not require speaker-specific training, and speaker-dependent,

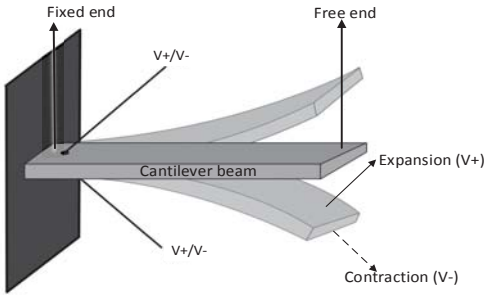


Fig. 1. A piezoelectric cantilevered beam working as VEH transducer.

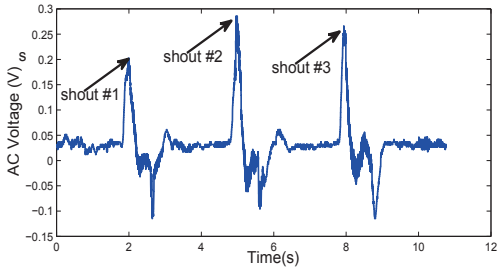


Fig. 2. Effect of shouting on the VEH piezoelectric beam.

which relies on speaker-specific training.

- We show that, in the direct scenario, VEH can detect hotwords with accuracies of 73% and 85%, respectively, for speaker-independent and speaker-dependent detections. We further demonstrate that these accuracies are comparable to what could be achieved with an accelerometer sampled at 200 Hz.
- Finally, for the direct scenario, we provide evidence that orientation of the piezoelectric beam relative to the speaker has a significant impact on hotword detection accuracy. This finding may serve as an important input to the design of next generation energy-harvesting mobile devices.

The rest of the paper is organized as follows. We provide a brief overview of VEH and its potential use as a hotword detector in Section II. VEH data collection process is explained in Section III, followed by hotword training and classification methods in Section IV. Results of VEH-based hotword detection are presented in Section V. Related work is reviewed in Section VI and the conclusions in Section VII.

II. VEH Overview

A. What is VEH?

Vibration energy harvesting (VEH) is the process of capturing environmental vibrations and converting them into electrical energy. Numerous vibration sources exist around us such as natural geographical vibrations (e.g. earthquakes), wind movement, machinery vibrations, human motion, and acoustic noise, to name a few. VEH has the potential to replace batteries or at least extending the battery life time for small, low-power electronic devices. Vibrations are typically converted into electrical energy using three transduction mechanisms [9]: piezoelectric, electromagnetic (capacitive), or electrostatic (inductive). Piezoelectric transducers are the most favourable due to their simplicity and compatibility with MEMS [10]. The piezoelectric effect was discovered in natural quartz crystals, but today's piezoelectric transducers are typically made from patented, proprietary ceramics.

Fig. 1 shows a typical usage configuration of a piezoelectric cantilevered beam to implement a VEH transducer. One end of the beam is fixed to the device, while the other is set free to oscillate (vibrate). When the piezoelectric material is subjected to a mechanical stress due to any source of environmental vibrations, it expands on one side and contracts on the other. Positive charges accumulate on the expanded side and negative charges on the contracted side, generating an AC voltage as the beam oscillates around the neutral position. The amount of voltage is proportional to the applied stress, which means that different vibration patterns would generate different AC voltage patterns. In most applications, the generated AC voltage is rectified to produce a DC, which can be used to power different sensors, such as an accelerometer, a gyroscope, or a microphone. However, the focus of our study is to investigate whether the AC signals can be used directly to detect hotwords.

B. Impact of speech on piezoelectric VEH

Human speech creates sound waves which move through the air in forms of air pressures. Therefore, a VEH should be able to detect changes in air pressures caused by human voice. To experimentally demonstrate this effect, we have asked a user to shout three times on top of a piezoelectric cantilever, while the generated voltage signal is being recorded. Fig. 2 shows the impact of the air pressure on the piezoelectric material. The device is responding by giving a voltage peak each time the air pressure hits the beam. This small experiment provides clear evidence that VEH can be used as a potential sensor to detect the presence of speech. Because the patterns of air pressures would be different when the human pronounces different phrases, we should be able to detect hotwords using VEH.

C. Study of VEH for hotword detection

Pervasive hotword detection requires continuous sensing of audio signals, which results in significant energy consumption when a microphone is used as an audio sensor. How to reduce audio sensing energy cost using other low-power sensors that can also register voice signals is a recent research trend in the literature. For example, researchers have shown that, instead of microphones, gyroscopes [8] or even accelerometers [7] can be used to detect hotwords at a fraction of the energy consumption. Assuming that the future wearable devices will have VEH to harvest energy from the ambient vibrations including human speech [11], [12], our proposal of using VEH

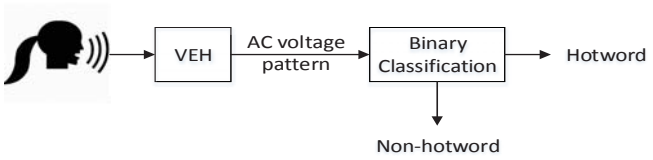


Fig. 3. VEH-based hotword detection.

output patterns for hotword detection will open the door for further power saving opportunities in wearable devices.

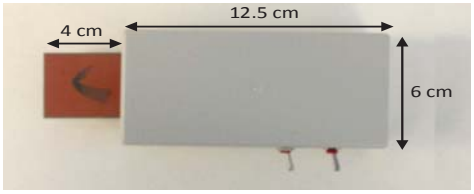
Fig. 3 shows the architecture used in our study for VEH-based hotword detection. The generated AC voltage data is continuously fed to a trained binary classifier, which classifies the input signal into either hotword or non-hotword. No actions will be taken during the normal conversation (speech contains no hotword), but if hotword is detected, the system will switch to the command mode. To realise the proposed binary classifier, we first need to collect AC data from both hotword and non-hotword speeches, and then train a suitable classifier to detect hotwords. These steps are explained in the following sections.

III. VEH Data Collection

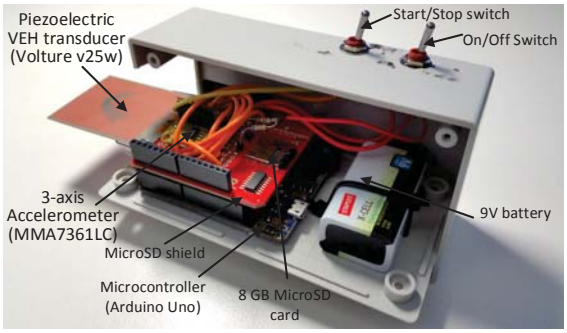
In this section, we explain our VEH hardware setup and the data collection process.

A. VEH Data Logger

We built a data logger to collect the output of a piezoelectric VEH. We chose a product called Volture from MIDÉ [13], which implements a piezoelectric VEH providing AC voltage as its output. In order to access the generated AC voltage signal of the VEH, an Arduino Uno is used as a micro-controller. We also included a 3-axis accelerometer (MMA7361LC) in the middle of the data logger, so any ambient vibrations can also be recorded in forms of accelerations. The Arduino micro-controller is programmed to sample the data from both the VEH and the accelerometer at a sampling rate of 1000 Hz. The sampled data is saved on an 8GB microSD card which is integrated to the Arduino using microSD shield. A 9V battery is used to power the Arduino. Two switches have been included, one to switch between the on/off mode of the device and the other to control the start and stop of data logging, which allows us to save AC data from different phrases into different files. Fig. 4 shows the internal and external appearances of our data logger.



(a) External Appearance.



(b) Internal Appearance.

Fig. 4. External and internal appearances of VEH data logger.

B. Experimental setup

We have collected data from many different experimental setups as summarised in Table I. We collected data from eight participants, four males and four females. Since our aim is to detect hotwords from phrases commonly used in typical conversations, we collect data in two different phases. In phase one, the user is asked to speak the hotword "OK Google" and repeat it 30 times. In the second phase, the user is asked to repeat each of three choices of a non-hotword phrases, "fine, thank you", "good morning", and "how are you" 10 times, giving a total of 30 'non-hotword' cases per user. The subjects are asked to utter all the phrases at their normal talking levels and take a break of few seconds between every two consecutive phrases. All experiments are carried out in a quiet room to eliminate background noise as much as possible.

C. VEH Usage Scenarios

During our data collection, we consider two possible usage scenarios of the VEH hardware, *direct vibrations* and *indirect vibrations*. The former scenario represents the case when the user is expected to bring the device close to his mouth when giving a command and talk directly on the surface on the piezoelectric beam. In our design, the piezoelectric energy harvester is left visible outside the VEH hardware case to implement the direct scenario (see Fig. 4(a)). The latter scenario is designed for cases when it is not practical or desirable to have a visible piezoelectric surface, but hotwords are expected to be detected from ambient vibrations captured by a VEH embedded somewhere in the mobile device. Data collection for these two scenarios are explained below.

- Direct vibration scenario:
  In this scenario, the user is asked to direct his voice towards the piezoelectric beam from a 3 cm distance. To study the impact of the piezoelectric beam's orientation on the hotword detection, we considered two different orientations. The data is first collected while the piezoelectric beam has a flat (horizontal) orientation. Then, the data is collected with the beam in its vertical orientation. Fig. 5 shows the two different orientations and how the direction of the airflow from user's speech affects the cantilever beam.
- Indirect vibrations scenario:

TABLE I
EXPERIMENTAL SETUP.

Participants	8 volunteers: 4 male and 4 female.
Classes	2 classes: 'hotword' and 'non-hotword': – 'hotword' class includes one phrase ∗ 'Okay Google' – 'non-hotword' class includes three phrases ∗ 'Fine, thank you'. ∗ 'Good morning'. ∗ 'How are you?'.
Dataset	60 instances/participant: 30 'hotwords' and 30 'non-hotwords'. In total, 480 instances.
Device orientation	2 orientations: horizontal and vertical (as shown in Fig.5).
Device position	on a table with 3 cm distance between subjects' mouth and the device.

In this scenario, the VEH is only expected to pick up the *ambient* vibrations caused by user speech in the vicinity of the device. The 3-axis accelerometer in the data logger is used to capture the ambient vibrations in terms of accelerations, which are later converted into VEH power signals using a second order mass spring damping model whereby the linear damper represents the combined damping offered by electrical and mechanical domains [4]. A second order mass spring damping model can be represented by a transfer function (in the Laplace domain) as in Equation (1).

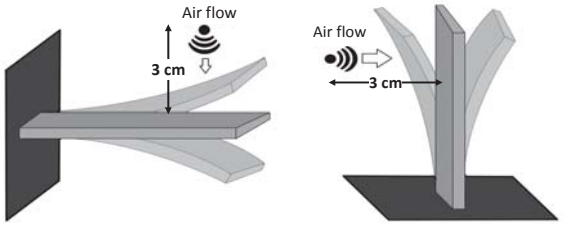
$$z(t) = \mathcal{L}^{-1}Z(s) = \frac{A(s)}{s^2 + \frac{b}{m}s + \frac{k}{m}}, \tag{1}$$

where m is the proof mass, k is the spring constant, b is the damping factor, $A(s)$ and $Z(s)$ denote, respectively, the Laplace transforms of the input force $a(t)$ and the proof mass displacement $z(t)$.

Simulink is used to simulate the response of this mass-spring-damper system. Once the gravity is filtered out from the data, the filtered data is converted to *proof mass displacement* using the previous Laplace domain transfer function. Next, the resulting proof mass displacement, $z(t)$, is limited by the limit of the proof mass displacement, Z_L. Finally, the generated harvested power is determined by:

$$p(t) = b\dot{z}^2(t) \tag{2}$$

This model has been used in [4], [14] to estimate the amount of power harvested from human motion vibrations. The configuration values, $m = 10^{-3}kg$, $Z_L = 10mm$, $k = 0.17$, and $b = 0.0005$, have been optimised in these research for typical human activities. In this study, we use the same model parameters to capture the vibrations generated by user speech in the vicinity of the device, as our main interest is hotword detection rather than power maximization and the VEH in mobile devices is likely to be configured to maximise power from human activity. The entire procedure is implemented using MATLAB and SIMULINK [15].



(a) Flat orientation. (b) Vertical orientation.

Fig. 5. Flat and vertical orientations of the VEH data logger.

Although, our prototype collects both VEH and accelerometer data at 1000 Hz sampling rate, most mobile devices restrict the accelerometer sampling rate to a maximum 200 Hz in order to reduce power consumption [7]. Therefore, we downsampled our accelerometer sampling rate to 200 Hz to match the current availability of accelerometer's sampling rate in mobile devices and to provide a fair comparison. In total, we have five different datasets, three of them are VEH datasets for indirect scenario and direct scenario for two different orientations (flat and vertical). The remaining two sets are accelerometer datasets for two different sampling rates 1000 Hz and 200 Hz. The accelerometer datasets are used for comparison purposes of both VEH and accelerometer-based hotword detection.

IV. HOTWORD TRAINING AND CLASSIFICATION

The VEH data obtained in both indirect and direct scenarios have been used to evaluate VEH-based hotword detection in comparisons with accelerometer-based hotword detection.

Feature extraction is a critical initial step in any classification process. This step transforms the input data into a set of features which are expected to extract the relevant information from the input data in order to perform the desired task. Table II shows our considered features set for VEH-based hotword detection, which was also considered previously in [7] for hotword detection from accelerometer data. The table shows single axis features, which are extracted from the single axis signals of VEH, in both direct and indirect scenarios, and each axis of the accelerometer signal separately. Besides the single axis features, the table shows multiaxis features which were extracted from the combination of the three axes of the accelerometer signal. Because all our hotwords were completed within 2 seconds, we used a time window of 2 seconds to extract the features.

For classification, we chose a Decision Tree (DT) classifier[1], which is a simple, yet powerful and very popular tree-based tool for classification and prediction [16]. The DT classifier is implemented in the widely used WEKA [17] software, which we used for our study. In the DT classifier, the classification process starts at the root of the tree and grows sequentially until reaching a leaf node. The central focus of the tree growing

[1]Authors in [7] also used DT

TABLE II
THE SELECTED FEATURES.

<table>
<tr><td>Time-domain features</td><td>
– Single axis features: calculated for VEH data and the three axes of the accelerometer x, y, and z separately:
∗ Mean: the central value of a window of samples
∗ Variance: a measure the amount of variation or dispersion from the mean.
∗ Standard Deviation: the square root of the variance.
∗ Minimum: the minimum value in a window of samples
∗ Maximum: the maximum value in a window of samples
∗ Range: The difference between the maximum and the minimum values in a window of samples
∗ Absolute Mean: average of absolute values,
∗ Coefficient of Variation: ratio of standard deviation and mean times 100; measure of signal dispersion,
∗ Skewness (3rd moment): measure of asymmetry of the probability distribution of the window of samples,
∗ Kurtosis (4th moment): measure of peakedness of the probability distribution of the window of samples,
∗ First, second and third quartiles: measures the overall distribution of the signal samples over the window,
∗ Inter Quartile Range: the difference between the upper (third) quartile and the lower (first) quartile of the window of samples; also measures the dispersion of the signal samples over the window,
∗ Mean Crossing Rate: measures the number of times the signal crosses the mean value; captures how often the signal varies during the time window,
∗ Absolute Area: the area under the absolute values of the signal samples. It is the sum of absolute values of the signal samples over the window,
– Multiaxis features: calculated as a combination of the three axes of the accelerometer:
∗ TotalAbsArea: sum of AbsArea of all three axis.
$$TotalAbsArea = \sum_{i=1}^{L} |Acc_x| + |Acc_y| + |Acc_z| \quad (3)$$
where $|Acc_x|$, $|Acc_x|$, and $|Acc_x|$ are the absolute values of the three axes of the accelerometer x, y, and z respectively. L is the length of the window.
∗ TotalSVM: the signal magnitude of all accelerometer signal of three axis averaged over the time window.
$$TotalSVM = \frac{\sum_{i=1}^{L} \sqrt{Acc_x^2 + Acc_y^2 + Acc_z^2}}{L} \quad (4)$$
</td></tr>
<tr><td>Frequency-domain features</td><td>
– Single axis features:: calculated for VEH data and the three axes of the accelerometer x, y, and z separately:
∗ DomFreqRatio: it is calculated as the ratio of highest magnitude FFT coefficient to sum of magnitude of all FFT coefficients.
∗ Energy: it is a measure of total energy in all frequencies. It is calculated as the sum of the squared discrete FFT component magnitudes.
$$Energy = \sum_{i=1}^{L/2} F_i^2 \quad (5)$$
where F_i is the magnitude of FFT coefficients.
∗ Entropy: captures the inpurity in the measured data. It is calculated as the information entropy of the normalized values of FFT coefficient magnitude.
$$Entropy = -\sum_{i=1}^{L} Fn_i log_2(Fn_i) \quad (6)$$
where Fn_i is the normalized value of FFT coefficient magnitude.
</td></tr>
</table>

algorithm is testing and selecting the feature with the most inhomogeneous class distribution, based on its information gain. The IG of feature f_i measures the expected reduction in entropy caused by partitioning the data (instances) according to this feature. The calculation of information gain is based on calculating the entropy H(S) of a set of classes S.

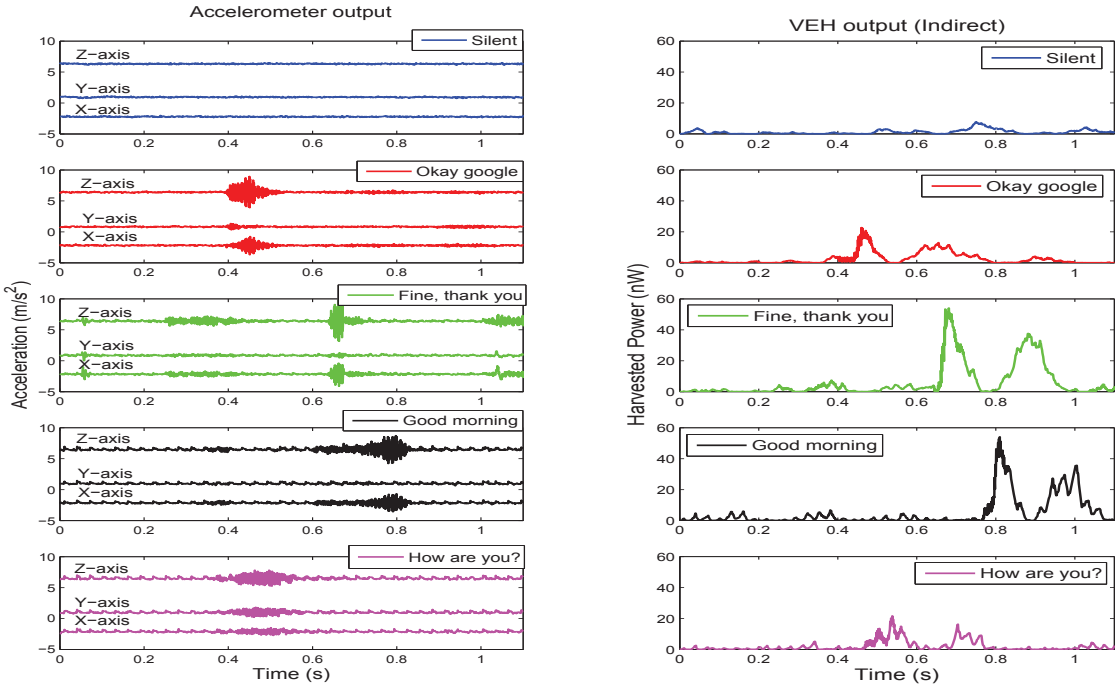
$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (7)$$

where n is the number of different activity classes and p_i is the proportion of all instances belonging to the i^{th} class. The information gain is then calculated using:

$$Gain(S, f_i) = H(S) - \sum_{v \in Values(f_i)} \frac{|S_v|}{|S|} H(S_v) \quad (8)$$

where S_v is the subset of S for which feature f_i has a value v (i.e., $S_v = s \in S | Values(f_i) = v$) and $|S|$ denotes the cardinality of the set S.

A well-known algorithm, which has been widely used for building decision trees over the years, is C4.5 [18]. In this algorithm, pruning is used to reduce the size of the tree to its optimal size, without reducing predictive accuracy. A tree that is too large, risks overfitting the training data and poorly generalizes to new samples. A small tree might not capture



(a) Raw acceleration data.

(b) VEH power estimated from acceleration data using the mass-spring model.

Fig. 6. Ambient vibrations captured by the internal accelerometer when different phrases were uttered by Female 1 compared to a silent case: (a) raw acceleration data and (b) VEH power estimated from acceleration data using the mass-spring model.

important structural information about the sample space [19].

In all usage scenarios, we evaluate two types of hotword detection, speaker-independent and speaker-dependent. In the speaker dependent case, the classification process is applied on the data collected from each individual participant. On the contrary, in the speaker independent hotword detection, all of the data gathered from the eight participants were first mixed and then fed to the classifier. In both cases, a 10-fold cross validation scheme [20] is used to get the results. In this scheme, the original data set is randomly divided into 10 equally sized subsets, where 9 of them are used for training and one subset is used for testing. This is repeated 10 times (the folds) and then the average of the results is reported.

V. Results

In this section, we present the results of our hotword detection study using piezoelectric VEH. We first present results for the indirect usage scenario, followed by the direct scenario. For both scenarios, we analyse the results for speaker independent as well as speaker dependent detections. The results of VEH-based hotword detection is compared with accelerometer-based detection. We also investigate speaker identification using piezoelectric VEH and compare it to the accelerometer-based identification. Finally, we analyse the impact of speech direction relative to the piezoelectric beam on the performance of hotword detection. In all of our results, we use the total accuracy as our evaluation metric. The total accuracy is calculated using Eq. 9 as a percent value.

$$Accuracy = \frac{TP + TN}{N} \times 100(\%), \tag{9}$$

where TP is the number of instances where speaking the hotword is correctly recognized as speaking the hotword, TN is the number of instances where speaking the non-hotword is correctly recognized as speaking the non-hotword, and N is the total number of instances.

A. Indirect Vibrations

Recall, that in this scenario, the VEH is expected to capture only the *ambient* vibrations caused by the user speech. Fig. 6(a) shows the 3-axial accelerometer output signals sampled at 1000 Hz, which represent these ambient vibrations. We see that there are no or negligible vibrations when the user remains silent (the top graph). However, the presence of ambient vibrations are clearly captured in the next four graphs. These results are in line with [7], which showed that human speech can be detected by accelerometers.

As explained in Section III, accelerometer traces can be used to estimate the power that can be potentially harvested by VEH. Fig. 6(b) shows estimated traces of power if VEH was used to harvest the ambient vibrations caused by user speech. We can see that the amount of power that could be

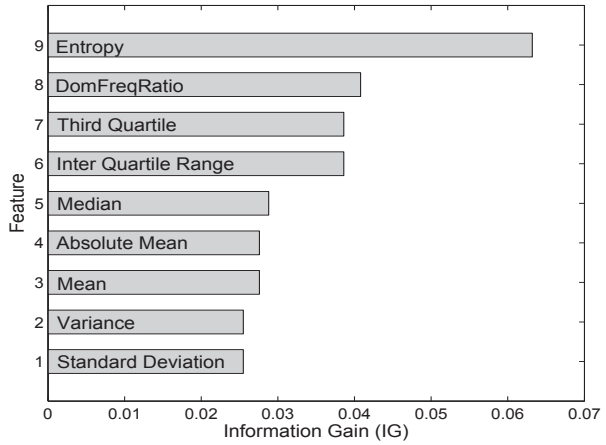


Fig. 7. Information gain of VEH power signals for the first nine features used for hotword detection.

harvested from such ambient vibrations is very low, in the order of tens of nW. However, we are not really interested in the amount of power generated by speech, but rather the patterns of power generation that could be used to detect hotwords. In this regard, we do see that the power amplitudes for all four phrases are certainly higher than the *silence* and that they all exhibit different power patterns. This implies that the indirect usage scenario may also be able to detect hotwords with some success.

To formally assess the discriminating capacity of the patterns of VEH power generation for hotword detection, we use the information gain theoretic analysis explained in Section IV. IG is a measure that determines how useful a given feature is for discriminating between the classes to be learned. Fig. 7 shows the IG of VEH power signals for the first nine features used for hotword detection. Indeed, this analysis of the features shows that many features provide positive gains, giving evidence that even these low power signals contain information to detect hotwords.

Table III shows the hotword detection accuracy results for indirect VEH. We find that for speaker independent, VEH can detect hotwords with 54% accuracy, which means that users would have to repeat the hotword just about once on average to get the voice control system into the command mode. However, the accuracy improved to 63% with speaker dependent training.

To see how these results compare to hotword detection using the 3-axial accelerometer itself, we conducted the training and classification with the acceleration data collected at 1000 Hz and sub-sampled at 200 Hz. Table IV shows the hotword detection results that could be achieved using the accelerometer. Once more, we found that *speaker dependent* outperform *speaker independent*, but we find that accelerometer achieves much higher accuracies than VEH. For example, even with 200 Hz, accelerometer can achieve accuracies of 76% and 87%, respectively, for speaker independent and speaker dependent

TABLE III
ACCURACIES (%) OF HOTWORD DETECTION FOR INDIRECT VEH.

Speaker Independent		**54.38**
Speaker Dependent	Female 1	53.33
	Female 2	63.33
	Female 3	46.67
	Female 4	56.67
	Male 1	78.33
	Male 2	66.67
	Male 3	70.00
	Male 4	75.00
	Average	**63.75**

TABLE IV
ACCURACIES (%) OF HOTWORD DETECTION FOR ACCELEROMETER.

Acceleometer Sampling Rate		200 Hz	1000 Hz
Speaker Independent		**76.04**	**83.13**
Speaker Dependent	Female 1	81.67	83.33
	Female 2	83.33	83.33
	Female 3	86.67	95
	Female 4	85	87.33
	Male 1	96.67	96.67
	Male 2	93.33	98.33
	Male 3	90	85
	Male 4	80	95
	Average	**87.08**	**90.5**

detections. The better performance of accelerometer compared to VEH can be explained by the 3-dimensional information available in the accelerometer (VEH has only 1-dimensional power data). However, VEH performance can be improved by harnessing voice vibrations more directly as examined in the following section.

B. Direct Vibrations

In this subsection, we examine the benefit of capturing voice vibrations more directly from the user. As explained in Section III, with this scenario, we conduct the training and classification using the AC voltage signal collected directly from the piezoelectric beam in our VEH data logger. Fig. 8 shows the patterns of AC voltage for silence and when the four phrases are spoken by Female 1. We see that voltage produced by silence is significantly lower than those produced by voice. We also notice that silence has a more periodic voltage pattern, which captures the background (noise) vibrations, while the voltage is markedly biased in the positive direction when phrases are spoken. This is expected because, in this scenario, sound waves continuously hit directly on one surface of the piezoelectric beam causing it to vibrate asymmetrically around the neutral position.

Table V presents accuracies when VEH AC Voltage is used for hotword detection with user speaking directly to a flat surface of the piezoelectric beam. Compared to the ambient vibration examined in the previous subsection, we can see marked improvement in the performance. With direct vibration capture, VEH can detect hotwords with accuracies of 73% and 85%, respectively, for speaker dependent and speaker independent detections, which are now comparable to accelerometer-based results with 200 Hz sampling.

TABLE V
ACCURACIES (%) OF HOTWORD DETECTION FOR DIRECT VEH.

Speaker Independent		**73.04**
Speaker Dependent	Female 1	81.67
	Female 2	76.67
	Female 3	88.33
	Female 4	88.33
	Male 1	96.67
	Male 2	85.00
	Male 3	75.00
	Male 4	93.33
	Average	**85.63**

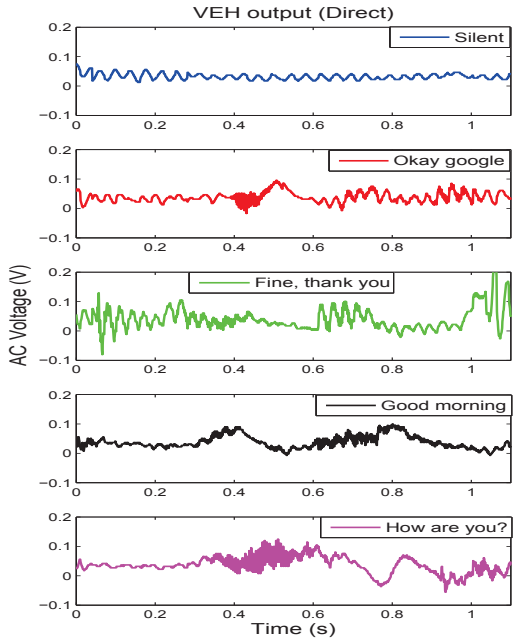


Fig. 8. VEH output signals for the direct scenario when different phrases were uttered by Female 1 compared to a silent case.

C. Speaker Identification

Previous work [7] has noted that accelerometer can be used to distinguish a user's voice from other users, which can be useful for user authentication applications. Therefore, in this subsection, we investigate VEH AC Voltage for speaker identification. To do so, we perform a multiclass classification by considering the data of each of the eight participants as a separate class. Fig. 9 shows the confusion matrix when VEH AC Voltage is used for speaker identification with user speaking directly to a flat surface of the piezoelectric beam. The results of the accelerometer-based identification are shown in parenthesis for comparison purpose. The results show that accelerometer outperforms VEH for speaker identification. The overall accuracy of VEH-based speaker identification is 56.87% compared to 85.83% for accelerometer-based identification. This reveals that VEH-based speaker identification still has a room for improvement which we consider as a future work.

Classified As

Actual User	F1	F2	F3	F4	M1	M2	M3	M4
F1	41 (50)	7 (0)	1 (0)	2 (0)	1 (6)	4 (3)	2 (0)	2 (1)
F2	8(1)	26 (55)	5 (2)	0 (1)	1 (0)	7 (0)	10 (0)	3 (1)
F3	4 (0)	6 (2)	29 (51)	6 (7)	2 (0)	1 (0)	9 (0)	3 (2)
F4	1 (0)	1 (3)	6 (4)	33 (49)	2 (0)	5 (3)	5 (1)	7 (0)
M1	2 (6)	0 (0)	2 (0)	2 (0)	47 (53)	0 (1)	7 (0)	0 (0)
M2	4 (2)	8 (0)	0 (1)	4 (5)	2 (0)	34 (45)	6 (4)	2 (3)
M3	2 (0)	3 (0)	4 (2)	4 (1)	7 (0)	3 (6)	34 (51)	3 (0)
M4	6 (0)	4 (0)	5 (0)	4 (0)	0 (1)	5 (1)	7 (0)	29 (58)

Fig. 9. Confusion matrix of the VEH-based speaker identification. Results of the accelerometer-based identification are shown in parenthesis for comparison purpose.

TABLE VI
ACCURACIES (%) OF HOTWORD DETECTION FOR VERTICALLY SPEAKING TO VEH.

Speaker Independent		**62.92**
Speaker Dependent	Female 1	88.33
	Female 2	80
	Female 3	83.33
	Female 4	65
	Male 1	90
	Male 2	80
	Male 3	56.67
	Male 4	83.33
	Average	**78.33**

D. Impact of VEH Orientation

Finally, we examine the impact of the orientation of the piezoelectric beam relative to the speaking or air flow direction. Table VI shows the accuracy results when the speaker is speaking vertically to the beam (see Section III). Interestingly, although the distance between the user and the beam is the same in both orientations, the vertical orientation degrades hotword detection performance significantly. These results show that if direct vibration usage scenario is planned for VEH-based hotword detection, VEH placement within the mobile device may have to be carefully designed.

VI. RELATED WORK

Voice control applications such as Siri [21] and Google Now [22] have emerged recently to improve user's interactivity. These voice control applications use hotwords such as "Okay Google" or "Hi Galaxy" to distinguish user's voice command from other conversations. One major challenge of voice control applications is the intensive sensing of audio signals which requires the microphone to be continuously ON to monitor user's voice commands [23]. One way to reduce the energy cost of audio sensing is the use of low-power sensors, e.g., accelerometer and gyroscope instead of the microphone.

MEMS sensors such as accelerometer and/or gyroscope have been widely used for human activity recognition [24]–[27] and indoor positioning applications. Matic et al, [28] have

also shown that accelerometer can be used for recognizing speech activity based on detecting phonation caused vibrations at the chest level. This can help in activating the voice control applications automatically, which usually require user interaction by a simple gestures on a button or using a Near Field Communication (NFC) tag.

In an attempt to reduce audio sensing energy cost, Michalevsky et al., [8] used the gyroscope sensor for digits recognition instead of the microphone. Gyroscope sensors consume less power than microphones, however, the authors in [8] had to upsample the received gyroscope samples at 4000 Hz to achieve acceptable accuracy, which is also power consuming. On the other hand, Zhang et al., [7] exploited the accelerometer sensor for energy-efficient hotword detection. They showed that accelerometer sampled at only 200 Hz can detect hotwords with comparable accuracy to microphones. They also showed experimentally that the accelerometer is more energy efficient than both microphone and gyroscope sensors.

To our knowledge, this is the first work to demonstrate that hotword detection is viable with VEH signals. Since VEH does not require power supply to generate AC signals, the possibility of VEH-based hotword detection opens up new power-saving opportunities for wearable devices.

VII. Conclusion and Future Work

To address battery issues, VEH is expected to be included in many emerging wearable devices. Using experiments with real subjects and VEH devices, we have shown that hotwords can be detected from the power generation patterns of VEH circuits with up to 85% accuracy. Given that VEH devices do not require any power supply to operate, VEH-based hotword detection has the potential for significant power saving. Our study has further revealed that the orientation of the VEH device relative to users' talking direction can have a major impact on the performance of VEH-based hotword detection. Future work will focus on experimenting with larger datasets, evaluating performance under user mobility and noisy environments, analysing power consumption of VEH-based hotword detection, and studying the impact of harvester size on the accuracy of hotword detection.

References

[1] S. P. Beeby, R. N. Torah, M. J. Tudor, P. Glynne-Jones, T. O'Donnell, C. R. Saha, and S. Roy, "A micro electromagnetic generator for vibration energy harvesting," *Journal of Micromechanics and Microengineering*, vol. 17, no. 7, 2007.

[2] R. Torah, P. Glynne-Jones, M. Tudor, T. O'Donnell, S. Roy, and S. Beeby, "Self-powered autonomous wireless sensor node using vibration energy harvesting," *Measurement Science and Technology*, vol. 19, no. 12, 2008.

[3] P. Mitcheson, E. Yeatman, G. Rao, A. Holmes, and T. Green, "Energy harvesting from human and machine motion for wireless electronic devices," *Proceedings of the IEEE*, vol. 96, no. 9, pp. 1457–1486, 2008.

[4] M. Gorlatova, J. Sarik, G. Grebla, M. Cong, I. Kymissis, and G. Zussman, "Movers and shakers: Kinetic energy harvesting for the internet of things," *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, no. 1, pp. 407–419, June 2014.

[5] K. Ylli, D. Hoffmann, A. Willmann, P. Becker, B. Folkmer, and Y. Manoli, "Energy harvesting from human motion: exploiting swing and shock excitations," *Smart Materials and Structures*, vol. 24, no. 2, p. 025029, 2015.

[6] S. Khalifa, M. Hassan, A. Seneviratne, and S. K. Das, "Energy-harvesting wearables for activity-aware services," *IEEE Internet Computing*, vol. 19, no. 5, pp. 8–16, Sept.-Oct. 2015.

[7] L. Zhang, P. H. Pathak, M. Wu, Y. Zhao, and P. Mohapatra, "Accelword: Energy efficient hotword detection through accelerometer," in *MobiSys15*, Florence, Italy, 18–22 May, 2015.

[8] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *23rd USENIX Security Symposium (USENIX Security 14)*, San Diego, CA, Aug. 2014.

[9] Y. Rao, S. Cheng, and D. P. Arnold, "An energy harvesting system for passively generating power from human activities," *Journal of Micromechanics and Microengineering*, vol. 23, no. 11, 2013.

[10] E. Lefeuvre, A. Badel, C. Richard, L. Petit, and D. Guyomar, "A comparison between several vibration-powered piezoelectric generators for standalone systems," *Sensors and Actuators*, vol. 126, no. 2, pp. 405–416, 2006.

[11] A. Seni, "Power of the Human Voice," http://large.stanford.edu/courses/2012/ph250/seni1/, submitted as coursework for PH250, Stanford University, Spring 2012.

[12] "Mobile phones could be charged by the power of speech," http://www.telegraph.co.uk/technology/news/8500161/Mobile-phones-could-be-charged-by-the-power-of-speech.html, accessed on 15 December, 2015.

[13] "Piezoelectric energy harvester," http://www.mide.com, accessed on 15 August, 2012.

[14] S. Khalifa, M. Hassan, and A. Seneviratne, "Pervasive self-powered human activity recognition without the accelerometer," in *Proc. International Conference on Pervasive Computing and Communication (PerCom)*, St. Louis, Missouri, USA, Mar 2015.

[15] S. Khalifa, "Inertial Kinetic Energy Harvesting Model," http://www.mathworks.com/matlabcentral/fileexchange/53585-inertial-kinetic-energy-harvesting-model.

[16] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 10, no. 1, pp. 119–128, 2006.

[17] "WEKA Software," http://www.cs.waikato.ac.nz/ml/weka/, accessed on 15 September, 2015.

[18] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1993.

[19] Y. Mansour, "Pessimistic decision tree pruning based on tree size," in *Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997.

[20] S. Kale, R. Kumar, and S. Vassilvitskii, "Cross-validation and mean-square stability," in *Proceedings of the 25th International Conference on Supercomputing*, Loews Ventana Canyon Resort, Tucson, Arizona, 31 May - 4 June 2011.

[21] "Apple siri," https://www.apple.com/ios/siri/, accessed on 15 September, 2015.

[22] "Google now," http://www.google.com/landing/now, accessed on 15 September, 2015.

[23] Y. Zhong, T. V. Raman, C. Burkhardt, F. Biadsy, and J. P. Bigham, "Justspeak: Enabling universal voice control on android," in *Proceedings of the 11th Web for All Conference*, ser. W4A '14, 2014.

[24] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Proceedings of the 17th conference on Innovative applications of artificial intelligence*, Pittsburgh, Pennsylvania, 9-13 July 2005.

[25] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2010.

[26] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Procedia Computer Science*, vol. 34, pp. 450 – 457, 2014.

[27] O. Lara and M. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications on Surveys and Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[28] A. Matic, V. Osmani, and O. Mayora, "Speech activity detection using accelerometer," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2012.