HiddenCode: Hidden Acoustic Signal Capture with Vibration Energy Harvesting

Guohao Lan, Dong Ma, Mahbub Hassan, and Wen Hu
University of New South Wales & Data61-CSIRO, Australia
{firstname.lastname}@unsw.edu.au

***Abstract*—The feasibility of using vibration energy harvesting (VEH) as an energy-efficient receiver for short-range acoustic data communication has been investigated recently. When data was encoded in acoustic signal within the energy harvesting frequency band and transmitted through a speaker, a VEH receiver was capable of decoding the data by processing the harvested energy signal. Although previous work created new opportunities for simultaneous energy harvesting and communication using the same hardware, the communication makes annoying sounds as the energy harvesting frequency band lies within the sensitive region of human auditory system. In this work, we present a novel modulation scheme to completely hide all communications within background music sound. The proposed modulation exploits sound masking theory to maximize signal to noise ratio of data communication without being audible to the music listener. We capitalize on the existence of repetitive sound patterns within popular music to realize synchronization between the transmitter and the receiver. We implement the proposed modulation within multiple hit songs and demonstrate its efficacy using a real VEH prototype made from off-the-shelf hardware. A user study involving 30 subjects confirms that the proposed modulation can completely hide VEH-based data communication from human perception while achieving up to 14 bps data rate, which is sufficient to transmit short codes or coupons of practical use.**

I. Introduction

The pervasive presence of speakers and microphones in consumer mobile devices offer the opportunity to transmit data using acoustic signal. Such audio-based communication has some distinct advantages over the conventional radio frequency (RF) communication interfaces, such as Near Field Communication (NFC), Bluetooth, and Wi-Fi. For example, the RF interfaces would either require lengthy configurations or may not be available (or turned on) in all devices. In contrast, the speakers and microphones are more pervasive and they are always ready to be used. Due to such advantages, audio-based communication has recently attracted significant attention in the research community [1], [2], [3].

Current solutions for audio-based communication rely on complex digital signal processing, such as the Fast Fourier Transform (FFT), to obtain frequency-domain information from the time-domain microphone samples, which consumes significant device power [2], [3]. To address the energy consumption problem in audio-based communication, we recently proposed [4] the use of vibration energy harvester (VEH) as an energy-efficient receiver for audio-based data communication. The significance of VEH lies in the fact that audio-based communication can be realized directly from the energy harvesting signals without engaging a microphone and the associated digital signal processing. As a matter of fact, mobile devices that can harvest energy from human motion to charge a standby battery have already started to appear in the market [5], [6], [7], [8]. These developments together with the proposed VEH-based audio communication techniques are fostering new pervasive computing research, which considers VEH not only as a source of ambient power for the sensor nodes [9], [10], but also as an instrument for mobile sensing and communication [4], [11], [12], [13], [14].

A key drawback of VEH-based audio communication is its highly *audibility* as it has to employ low frequencies, e.g., between 50 Hz to 200 Hz, to harvest energy from typical environmental vibrations. As high audibility is likely to annoy most users, techniques must be developed to address the audibility problem if VEH is to be used for data communications. However, solving the audibility problem in this low frequency band is a formidable challenge for a number of reasons. VEH has a very narrow resonance bandwidth on the order of a few Hz only [15], [16]. This means that making VEH communication more *melodic* (less annoying) is not possible because melody is constructed by varying the sound frequencies over time within a wide bandwidth to distinguish different musical notes (different notes encode different data symbols) [1], [17]. Masking audio communications within music is another option explored by other researchers [2], [3], but they employ very high-frequency acoustic signals between 16-20 kHz, which are easy to mask with low frequency music sound.

Nevertheless, music-based masking is an attractive option available to address the acute audibility problem of VEH-based audio communication. If we can successfully hide the VEH communication within background music, it can be applied in many pervasive applications such as delivering coupons to shoppers, providing additional information from TV to the mobile device, and so on. In this work, we therefore undertake a detailed study to realize musical masking of VEH-based audio communication. To the best of our knowledge, this is the first work attempting to capture hidden acoustic data within music using VEH. The contributions of this paper can be summarized as follows:

- Using 30 hit songs and a real VEH prototype, we analyze the music power available within the narrow resonance band of the VEH hardware. Our analysis reveals that,

for majority of the data symbol intervals, original music power is too weak to vibrate the energy harvester adequately for accurate symbol detection. Fortunately, a large fraction of these affected symbols can be recovered by applying the theory of *frequency masking*, which allows us to raise the transmission power up to a *masking threshold* without interfering with the music. Nevertheless, the low transmit power problem arising from the extremely narrow VEH resonance frequency band entails that data can be encoded only in high-energy symbol intervals where the pattern of these data symbol intervals are dictated by the host music, but must be communicated to the VEH receiver.

- We propose to exploit *repetitions* prevalent in popular songs to seamlessly communicate the data interval pattern to the VEH receiver. We demonstrate that the VEH can self-learn the *good* (high energy) and *bad* (low energy) intervals by simply 'listening' to the first part of the repeated music. It can then use the observed symbol pattern to identify the data intervals and extract the data bits from the repeated music, which has the same pattern.
- We implement the proposed repetition-based encoding in four hit songs, conduct subjective tests for audibility, and analyze bit error rate (BER) for the decoder using a real VEH device. Based on the mean opinion score (MOS) obtained from 30 subjects, the proposed masking approach can achieve up to 14bps without annoying the user.

The rest of the paper is organized as follows. We review the related work in Section II. Auditory masking preliminaries are presented in Section III. Analysis of hit musics is presented in Section IV. The proposed repetition-based data encoding and decoding are described in Section V followed by its performance evaluation in Section VI. We discuss relevant ongoing advancements in VEH technology in Section VII before concluding the paper in Section VIII.

II. Related Work

A. Audible Data Transmission

In sound based data communication, the audibility of sound signal is a double-edged sword. On one hand, it enables the users to be aware of the existence of data transmission, on the other hand, it is annoying to the users in case of long data transmission. To address this problem, Lopes et al. proposed to synthesize different musical instruments to make the acoustic signal pleasant [1]. In their work, they modulate data information in piano, bell, and clarinet sounds, and encode the data into musical notes. A M-ary frequency shift keying (MFSK) modulation scheme is designed for short range communication. A similar work has been proposed by Madhavapeddy et al. [17], in which they design a melodic based method for data transmission. A melody is constructed by varying the frequencies of the sound over time based on a pre-defined sequence which is known to both transmitter and receiver in advance. Their prototype system is able to achieve a bit rate of 20bps over a short distance. However, none of those solutions can be adopted for VEH-based acoustic communication, as VEH has a very narrow resonance frequency band [15] which makes it impossible to leverage the multi-frequency based schemes.

B. Inaudible Data Transmission

The current trend in the literature is to push the data communication to the high-frequency band (i.e., above 18kHz). There are two major advantages of using high-frequency band for communication. First, a sound with a higher frequency is able to be transmitted at a high energy level without perception of the human auditory system [18], [19]. Second, according to the masking theory, higher frequency sounds are much easier to be masked than low frequency sounds [20]. Thus, for high-frequency sound transmission, the embedded data can easily be hidden from the user, by leveraging the masking effect of human hearing system. Recent works, such as [2] and [3], all utilize the masking effect to hide the embedded acoustic signal from human ears. The Dolphin system in [2], is able to transmit at 500bps within one meter by adopting the OFDM modulation scheme. In [21], the authors proposed an ultrasound chirp-based communication system using commercial speakers. The system is able to transmit 16bps data up to 25m distance. A more recent work in [22], the authors developed a near-ultrasound chirp-based system as the second screen service between TVs and smart devices. To improve the data rate, a chirp quaternary orthogonal keying has been proposed. However, as VEH only response to the acoustic signal in the low frequency range (i.e., below 500 Hz), none of the above systems can be used for VEH-based acoustic data communication.

III. Preliminaries on Auditory Masking

The human auditory system is able to capture the sound with frequency ranges from 20 Hz to 20 kHz. There is a *threshold in quiet* (Figure 1), which indicates, as a function of frequency, the minimum sound pressure level of a pure tone that the human auditory system is able to hear with no other sound present. As shown in Figure 1, the human ear is most sensitive to sound with frequency between 20 to 18KHz [19], but is almost deaf to the sound with frequency either below 20Hz (infrasound) or above 20KHz (ultrasound). Clearly, VEH-based audio communication, which works in the low frequency range below 200 Hz, can be easily heard by human even if the sound pressure level is very low. On the other hand, audio communications in the near ultrasound frequencies cannot be easily heard even if the sound pressure level is high. Because of this benefit, most of works [2], [3] use near ultrasound frequencies for device-to-device audio communication .

It is possible to mask a low-energy sound (maskee) in one frequency by a high-energy sound (masker) in another frequency [18]. In essence, the masker generates a *masking threshold* that elevates the *threshold in quiet* illustrated earlier in Figure 1. Any sounds or frequency components that have

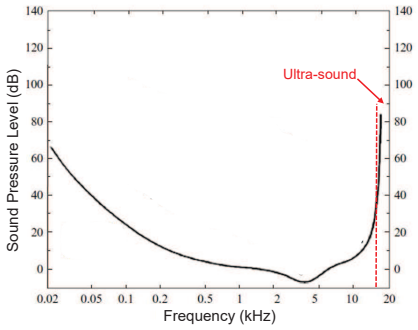


Fig. 1: The threshold in quiet [18].

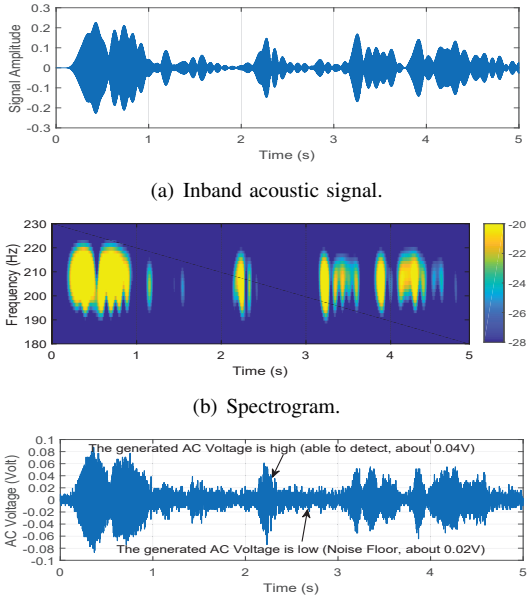
signal energy level below this *masking threshold* is masked by the presence of the masker [20]. Such auditory masking is frequently used in audio watermarking [20], [23], [24] to hide digital data, such as copyright information, into the audio.

Psychoacoustic model I [23] is an efficient algorithm that can be used to estimate the *masking threshold* generated by a given masker. The algorithm consists of four steps [20], [23]. First, the input audio signal (masker) is analysed to identify tonal (sinusoidal) and non-tonal (noisy) components in the audio, as their masking models are different. Second, spreading functions are used to mimic the excitation patterns of both tonal and non-tonal maskers. Third, after removing the maskers with energy below the absolute hearing threshold and the tonal components separated by less than 0.5 Barks scale, a *global masking threshold* is estimated. Finally, the masking threshold is obtained as the lowest level of the global masking threshold in the 32 equal width sub-frequency bands of the entire spectrum. For more details about the algorithm, readers are referred to [20], [23].

IV. Analysis of Music Energy

In any form of wireless communication, for a given distance between the transmitter and the receiver, the *transmit power* influences the signal-to-noise ratio (SNR) at the receiver and hence the bit error rate (BER) and data rate in bits per second (bps). The higher the transmit power, the better the SNR, and vice versa. If the transmit power is too weak, it may not be possible to decode the bits at the receiver at all (SNR too low). For VEH-based audio communication, the music power allocated to the VEH resonance frequency band acts as the transmit power of the acoustic communication link. If the music power in the VEH band is too weak, it may not be possible for the VEH to vibrate and decode data. We therefore set out to analyse the music power available in the VEH band as our first priority.

We have downloaded 30 hit songs from the Billboard Top 200 list [25]. We use the V25W piezoelectric energy harvester from MIDE [26] as the target VEH device which has a resonance band from 195 Hz to 210 Hz (half-power bandwidth). This means that the VEH can only respond to frequencies within this narrow (15 Hz) bandwidth. We therefore, analyse music power within this band by applying bandpass and bandstop filters to divide the original music



(a) Inband acoustic signal.

(b) Spectrogram.

(c) The AC voltage generated by the VEH receiver.

Fig. 2: Example of a five-seconds inband signal and the corresponding AC voltage generated by the VEH receiver.

signal into two parts: the *inband signal*, i.e., the signal within the VEH resonance band, and the *outband signal*, i.e., the signal outside the VEH resonance band. The inband signal is used as the *host signal* for data modulation, while the outband signal can be used to estimate the *masking threshold*, which enables us to raise the transmit power of the inland signal without making it audible.

For one of the songs, Figure 2 shows the energy of a five-seconds inband signal and the corresponding AC voltage generated by the VEH receiver when the music signal is transmitted through a speaker at a line-of-sight distance of 5cm. We can clearly observe that, for high energy music periods, the generated AC voltage is high and can be easily distinguished from the noise floor (data can be decoded), while it is too low to be separated from noise in other intervals (data cannot be decoded). This result suggests that data cannot be transmitted at any time. For the transmitted signal to be detected at the VEH, data must be encoded within the music intervals that have high inband energy.

In data communication, the fundamental time interval that carries data is called a *symbol* interval. Each symbol interval carries a single symbol, but a symbol can carry one or more bits depending on the modulation used. For binary modulations, such as the vibration on-off keying (VOOK) used for VEH-based audio communications in [4], each symbol carries a single bit, a '0' or '1'. By analysing the inband signal energy at the resolution of symbol intervals, we can therefore work out how many symbols and data bits we can transmit per second using a particular music as a host. Obviously, we will only count the symbol intervals, which have high energy, and discard the others as unusable (transmit power is too weak).

Let us define the symbol counting and data rate calculation

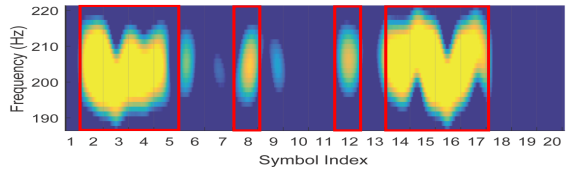
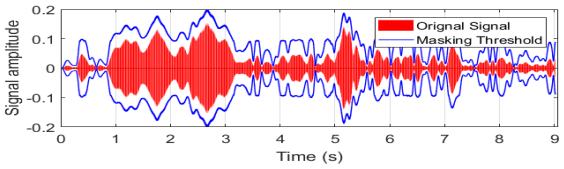
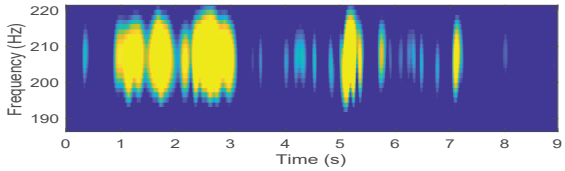


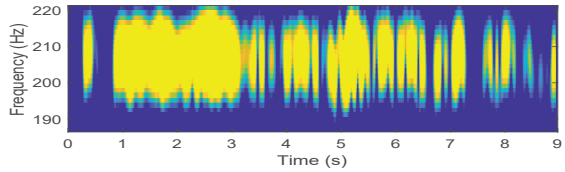
Fig. 3: The spectrogram of a 2-sec music trace. The symbol intervals with high energy are highlighted in red rectangles.



(a) Signal amplitude.



(b) Spectrogram of original signal.



(c) Spectrogram of amplified signal.

Fig. 4: Example of a nine-seconds signal. (a) compares the amplitude of the original inband music signal with the estimated masking threshold; (b) the spectrogram of the original inband music signal; and (c) that of the amplified signal.

process more precisely. For a music trace of T seconds, we have $n_{symbol} = \frac{T}{t_{symbol}}$ symbol intervals, in which t_{symbol} is the symbol interval length. A symbol interval will be regarded as a *data symbol* if β percentage of the symbol interval has inband energy higher than a pre-defined threshold, E_{thres} (-22dB in our setting). Obviously, more data can be sent for smaller β, but at the risk of higher BER. Extensive experiments reveal (details are in Section VI) that for 100 ms symbol intervals, BER can be kept below 2% for $\beta \geq 0.3$.

Figure 3 shows an example of the data symbol intervals for a 2-sec music trace using $t_{symbol} = 100$ms and $\beta = 0.5$. In this case, out of 20 symbol intervals, only 10 intervals can be used for transmitting data symbols. For a binary modulation (one bit per symbol), this would allow transmission of 10 bits within 2 seconds, yielding an effective data rate of 5 bps.

Recall that it is possible to raise the transmit power of the inband signal up to the *masking threshold* without making it audible. We employ the Psychoacoustic model I explained in Section III to estimate the masking threshold of the inband signal as shown in Figure 4. We can clearly observe that

TABLE I: Percentage of symbol intervals that can be used for data transmission averaged over the 30 music traces.

Symbol Length	**Method**	β		
		0.3	0.4	0.5
50ms	Non-Masking	19%	18%	14%
	Masking	56%	54%	51%
100ms	Non-Masking	21%	19%	15%
	Masking	57%	55%	51%

the use of *audio masking* has made more symbol intervals to be feasible for data transmission. For the 30 hit songs, Table I compares the percentage of symbols available for data transmission using masking (i.e., Masking) against that without using masking (i.e., Non-Masking) for two different symbol lengths. We can see that data encoding chance can be increased by a factor of 3 by masking the inband signal with the outbound music.

While masking can help increase the number of usable symbol intervals in the music for data transmission, it is still not possible to encode data in *every* symbol. This creates a new problem that has not been faced by previous works [2], [3], which were able to transmit in every symbol thanks to the high transmit power allowed in the near ultra-sound frequencies. When every symbol is used for data, all symbols in the data packet are used for encoding and decoding. On the other hand, when only specific symbols can be used for data encoding, some symbols in the data packet do not contain any information and must be ignored when decoding. The problem is that unless the decoder knows the data symbol patterns used by the encoder, it cannot perform decoding successfully. Masking VEH-based audio communication with music therefore requires a solution to this new problem. In the following section, we propose a novel encoding and decoding technique that solves this problem by exploiting the prevalent repetition of sounds in popular musics.

V. Repetition-based Encoding and Decoding

There are several possible ways for the VEH receiver (decoder) to learn the data symbol pattern used by the the transmitter (encoder). A naive way is to publicly share the data symbol patterns, so that the demodulator is aware of the patterns in advance. However, this would require the demodulator to store data symbol patterns for a large number of music and it needs to be trained to detect the music. Another option could be to use RF channels, such as Bluetooth or WiFi to transfer the symbol patterns before the music is played, but that would perhaps consume more energy than that would be saved by using the VEH. We propose a novel encoding technique that exploits the prevalent repetitions within popular songs to let the VEH self-learn the symbol pattern.

A. Repetitions in Music

To appreciate the proposed encoding, we must understand how music repetitions work. Repetition in music refers to the phenomenon where a sequence of sound is repeated a few times. Repetition plays a very important role in all kinds of music. It makes the music easier to be remembered by the

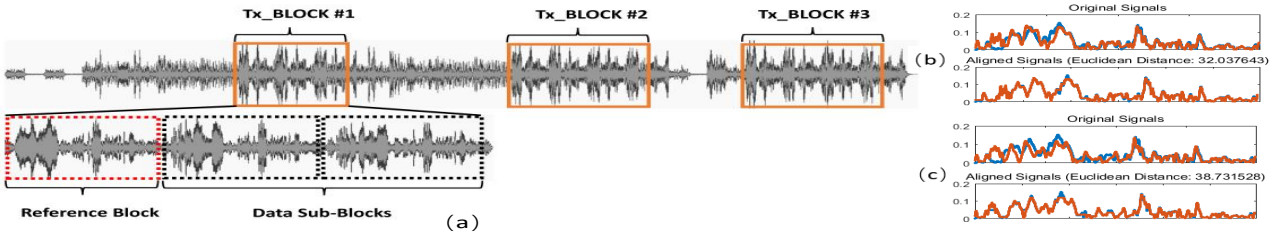


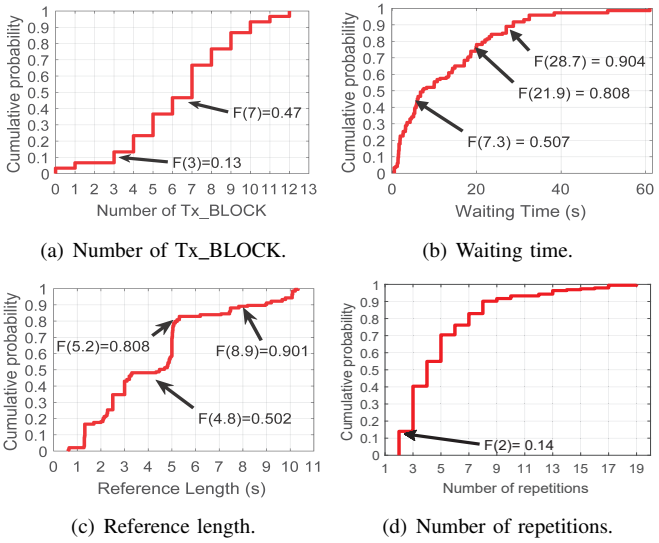
Fig. 5: (a) example of the music repetitions and repetition-based encoding; (b) and (c) give the comparison between the Reference and Data Sub-Blocks in terms of the original signals and the aligned signals obtained from DTW.

listeners [27], [28]. Repetition is extraordinarily prevalent and has been extensively used since the early stage of music [29]. As an example of repetitions, Figure 5(a) plots the inband signal for a music trace (i.e., "Dreams" by Nana). We can observe that, within the entire music trace, there are three segments (highlighted in the solid rectangles) in which a same signal pattern has repeated several times. In addition, within each of those segments, there are a number of repetitive signal periods (i.e., repetitions that highlighted in black dotted rectangles). Figure 5(b) and (c) compare the signal in the Reference Block with the signal in the two subsequent Data Sub-Blocks, respectively. We can observe the similarity between those repetitive signal periods from the distance obtained from Dynamic Time Warping (DTW) [30].

In the following, we conduct a statistic analysis based on a dataset of 30 musics to investigate the repetitive pattern in the response-band of VEH. First, we define the following terminologies:

- *Signal Repetition*: indicates the piece of acoustic signal that has repeated several times.
- *Transmission Block (Tx_BLOCK)*: the signal segment in the music trace which contains several signal repetitions. Figure 5 highlights three Tx_BLOCKs in orange rectangles that appear within the music trace.
- *Reference Sub-Block*: the *first* signal repetition that appears in the Tx_BLOCK. It serves as the 'location key' to inform the demodulator the locations of data intervals (symbol pattern) in the upcoming data payload.
- *Data Sub-Block*: is the actual signal window in which data bits will be embedded in. As shown in Figure 5, the Data Sub-Blocks starts from the end of the Reference Sub-Block, and includes all the remaining signal repetitions in the Transmission Block.

Specifically, we are interested in the following parameters: (1) the number of Transmission Block appearing in a music trace; (2) the transmission waiting time between any two adjacent Transmission Blocks; (3) the length of Reference Sub-Block inside each Transmission Block; and (4) the number of repetitions appear in a single Transmission Block. The cumulative distribution function (CDF) of those parameters are shown in Figure 6. Figure 6(a) indicates that 87% of the examined music traces have more than three Tx_BLOCKs within the entire signal trace. Moreover, as shown in Figure 6(b), given $F(28.7) = 0.904$, we can conclude that 90.4% of the musics



(a) Number of Tx_BLOCK. (b) Waiting time.

(c) Reference length. (d) Number of repetitions.

Fig. 6: The CDF of different parameters.

have a transmission waiting time less than 28.7 seconds. Figure 6(c) gives the CDF of the reference length. As we are expecting the Reference Sub-Block to be as short as possible, such that more time in the Tx_BLOCK can be used in the data payload for data transmission. Based on the 193 Tx_BLOCKs we have extracted from the 30 music traces, we can observe that 90.1% of the examined transmission opportunities have an reference less than 8.9 seconds. Lastly, Figure 6(d) gives the CDF of the number of repetitions appear in each of the 193 Tx_BLOCKs. As shown, given $F(2) = 0.14$, we can conclude that 86% of the Tx_BLOCK include at least three signal repetitions.

B. Encoding Design

In the following, we discuss the design of encoding. First, the encoder needs to prepare the signal before data encoding. It copies the signal in Reference Sub-Block to replace the original signal in the Data Sub-Block. This makes all the repetitions have the same signal pattern before encoding, such that, the locations of the data intervals in the Data Sub-Blocks should be the same as that in the Reference Sub-Block. During encoding, the signal in the Reference Sub-Block will remain unchanged as the 'location key', whereas, the data intervals in the Data Sub-Blocks will be modulated to encode data.

After preparing the signal, binary bits data are encoded into the Data Sub-Blocks by modulating the signal of the data

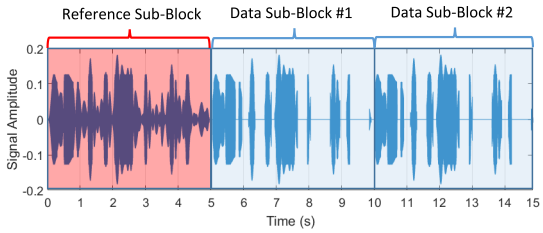


Fig. 7: An example of data encoding with repetitions.

intervals in an ON-OFF manner. For a symbol interval in the Data Sub-Blocks with index k, the modulated signal, $s'(k)$, is obtained by modifying the original signal, $s(k)$, as follows:

$$s'(k) = \begin{cases} s(k) & \text{if } k \text{ is a data symbol to encode data bit '1',} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Data bit '1' and '0' is encoded by maintaining and wiping out the original signal in the corresponding data interval, respectively. In addition, if symbol interval k is not a data interval, we wipe out its signal as well, to eliminate its interference on adjacent intervals. Our subjective perception study in Section VI-C proves that our encoding method has no adverse effect on the listening of the host music. As an example, Figure 7 shows a 15-seconds audio signal contains three repetitions. The first repetition is the Reference Sub-Block. No modification has been applied to the signal in it. The remaining two repetitions are the Data Sub-Blocks, for which the signal is originally copied from the Reference, but has been modulated to encode data '1' or '0'. We can observe from Figure 7 that, comparing to the signal in the Reference, some parts of the signal in the Data Sub-Blocks have been wiped out, only the intervals that encoded with '1' are maintained.

C. Demodulation Design

Now, we shift focus to the design of demodulation. To accurately decode the modulated data, the demodulator needs to obtain the locations of the data intervals (i.e., 'location key') using the AC voltage signal harvested from the Reference Sub-Block. As an example, Figure 8 exhibits the AC voltage signal (the positive part only) generated by the VEH receiver from the sound signal in the Reference Sub-Block (shown in Figure 3). To obtain the 'location key', a voltage threshold, V_{thres}, is used. For a given symbol interval, if more than β percentage of time within which the generated AC voltage is higher than V_{thres}, the symbol interval will be regarded as data interval. Intuitively, as the data intervals in the Reference Sub-Block contain high energy inband signal during the modulation (as shown in Figure 3), the transmitted sound will generate high AC voltage when it received by the VEH receiver. As demonstrated in Figure 8, by using the thresholding method, the receiver can obtain the 'location key' from the AC voltage signal. The detected data intervals are highlighted in green rectangles, which match with the data intervals used by the modulator (shown in Figure 3).

After obtaining the locations of data intervals, demodulator starts to decode the embedded data in the upcoming Data Sub-Blocks by analyzing the gradient (i.e., slope) of the voltage

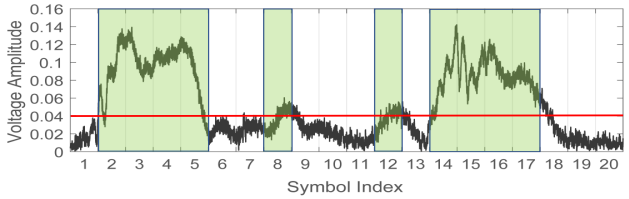


Fig. 8: Detecting the data intervals from the AC voltage signal.

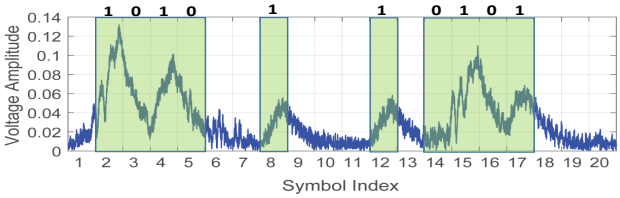


Fig. 9: Decoding the embedded binary bits by analyzing the AC voltage signal at the data intervals.

signal in the corresponding data intervals. For a given data interval with symbol index k, the embedded binary data, $d(k)$, is determined by analyzing the numerical gradient of the AC signal, $ac(k)$, in the corresponding symbol interval k:

$$d(k) = \begin{cases} 1 & \text{if there exhibits a positive gradient in } ac(k), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Intuitively, as defined in Equation 1, binary bit '1' is represented by maintaining the high energy signal in the data interval, whereas, bit '0' is represented by wiping out the original signal. Consequently, the data intervals that encoded with '1' will make the VEH receiver vibrating and the AC voltage generated by the VEH will increase. In contrast, data intervals that embedded with '0' will make the VEH stop vibrating and the harvested AC voltage will decrease (or maintain low if it was low in the previous interval). Therefore, by simply looking at the gradient of the AC voltage, the VEH demodulator can correctly decode the embedded '1' and '0'.

As shown in Figure 8, the demodulator detects from the Reference Sub-Block that symbol intervals with index equals to 2, 3, 4, 5, 8, 12, 14, 15, 16, 17, and 18, are the data intervals. Then, to decode the embedded data, the demodulator starts to analysis the AC voltage in the corresponding data intervals. As an example, Figure 9 plots the AC voltage signal generated by the VEH from a data string of '1010110101'. We can notice that the AC voltage signal has a positive slope in symbol interval 2, 4, 8, 12, 15, and 17. In contrast, the AC voltage is decreasing or maintaining low in symbol interval 3, 5, 14, and 16. Thus, following the decoding policy in Equation 2, the demodulator can successfully decode the embedded data bits as '1010110101'.

VI. System Evaluation

A. Prototype Implementation

Figure 10 shows our VEH prototype. The off-the-shelf V25W piezoelectric vibration energy harvester from MIDE [26] is used to build our receiver. The V25W energy harvester is mounted on a clamp base with its output bins

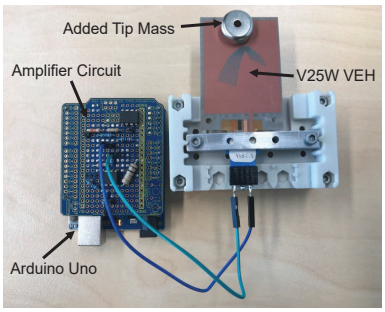


Fig. 10: The VEH receiver prototype.

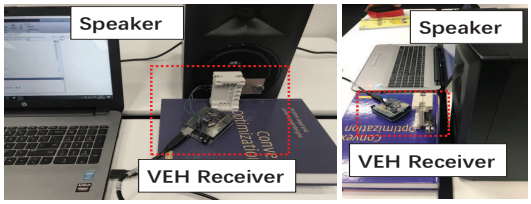


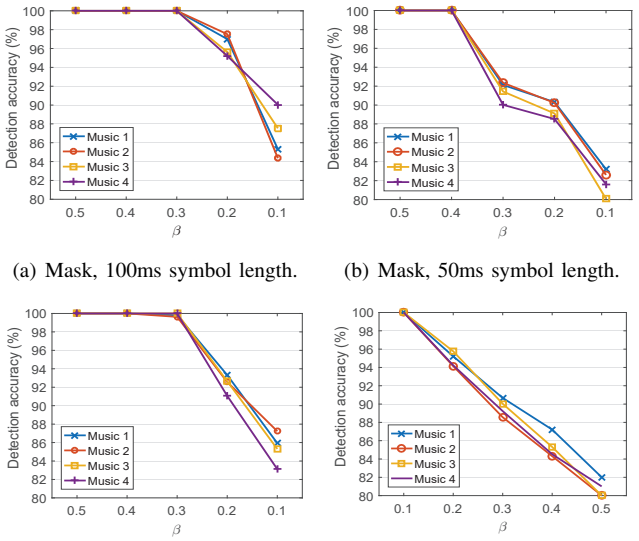
Fig. 11: Experiment setting (views from different angles).

connected to an amplifying circuit. An Arduino UNO board is used as the host platform to sample the amplified AC voltage output from the amplifying circuit through the onboard 10-bit analog-to-digital converter (ADC) at 1KHz frequency. The sampled AC voltage data are stored for further analysis. As we are only interested in the harvested voltage signal for decoding, we do not physically store the generated power for energy harvesting purpose, only the output AC voltage from the harvester is sampled and stored. However, simultaneous energy harvesting and communication is able to achieve by utilizing an specially designed circuit [11], [31].

B. Decoding Performance

In the following, we evaluate the decoding performance. We consider four musics: (1) "Closer" by Lemaitre; (2) "Dreams" by Nana; (3) "Lines & Circles" by O-Town; and (4) "Shape Of You" by Ed Sheeran. We extract a 10 to 20 seconds signal segment from each of the musics to be used as the *Original* trace. The extracted segment corresponds to a Tx_BLOCK in the music signal. Then, for each of the four original traces, we encode a string of binary data bits in it (i.e., a bit string of '101010...1010' in which '1' and '0' appears alternatively). Figure 11 presents the experiment setting. The modulated audio traces are played through an external speaker (i.e., JBL LSR305 loudspeaker) connected to the laptop. As the speaker causes the table to vibrate when playing out sound, we place the VEH receiver on a different table to eliminate the effect of such mechanical vibrations. During the experiment, the distance between the VEH receiver is fixed to 5cm, and the volume of the speaker is configured to 70dB SPL (which is comparable to typical conversation [32]).

We evaluate the performance of our system in two different settings, denoted as HiddenCode-Mask and HiddenCode-NonMask, respectively. In HiddenCode-Mask, the psychoacoustic model is applied to improve the inband signal energy and data encoding chance. In contrast, in HiddenCode-



(a) Mask, 100ms symbol length. (b) Mask, 50ms symbol length.

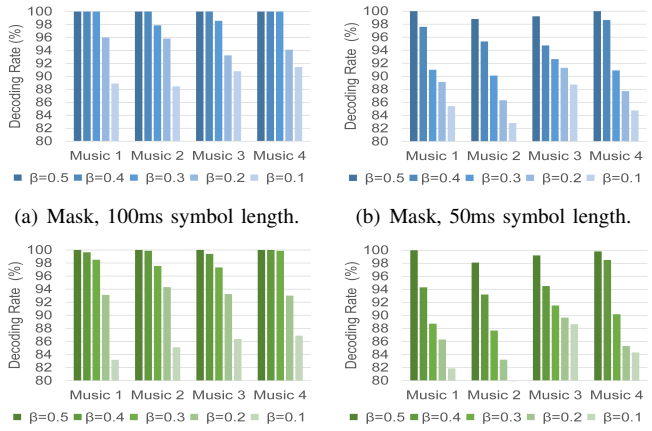
(c) NonMask, 100ms symbol length. (d) NonMask, 50ms symbol length.

Fig. 12: The achieved accuracy in data interval detection.

NonMask, the original music signal is used without applying frequemcy masking. Two different symbol interval lengths are considered, 100ms and 50ms, and for each of the symbol lengths, we adjust the value of β from 0.1 to 0.5, to investigate its impact on the decoding performance.

*1) **Accuracy in Data Interval Detection**:* First, we examine the accuracy achieved by the decoder in detecting the data intervals in the Reference Sub-Block by using the harvested AC voltage signal. Figure 12 exhibits the detection accuracy as a function of β and symbol length for the four music traces. These figures clearly demonstrate that, for a given music and symbol length, higher detection accuracy can be achieved by (1) using a larger β, and (2) by applying the frequency masking.

Recall our discussion in Section IV that a larger β is able to filter out some low energy symbol intervals from the selection of data intervals, only those contain high energy signal are remained. As a result, a larger β enables better detection accuracy at the VEH receiver, and vice versa. Recall the example shown in Figure 3 and Figure 8, with $\beta = 0.5$, the receiver can accurately detect those high energy data intervals from the AC voltage signal with 100% accuracy. However, with a smaller β, some low energy symbol intervals will be selected as data intervals by the modulator. For instance, with $\beta = 0.2$, the symbol interval 6 and 9 shown in Figure 3 will be selected for data transmission. Unfortunately, as shown in Figure 8, the AC voltage generated in symbol interval 6 and 9 is below the voltage threshold, and thus, receiver fails to detect those two data intervals and results in detection error. In contrast, we can benefit from the use of frequency masking, as it can raise the signal energy of the data intervals up to the masking threshold. Consequently, at the VEH receiver side, the generated AC voltage from the data intervals will become more distinct from the voltage threshold.



(a) Mask, 100ms symbol length. (b) Mask, 50ms symbol length.

(c) NonMask, 100ms symbol length. (d) NonMask, 50ms symbol length.

Fig. 13: The achieved decoding rate with different settings.

TABLE II: The encoding rate of HiddenCode-Mask for different music traces given different symbol lengths.

Trace	Symbol Length	β				
		0.1	0.2	0.3	0.4	0.5
Music 1	50ms	46%	44%	43%	40%	38%
	100ms	52%	48%	46%	42%	38%
Music 2	50ms	50%	48%	46%	43%	42%
	100ms	52%	48%	47%	45%	45%
Music 3	50ms	79%	77%	76%	74%	71%
	100ms	84%	82%	78%	76%	76%
Music 4	50ms	60%	58%	56%	52%	49%
	100ms	71%	69%	65%	53%	49%

*2) **Decoding Rate & Effective Data Rate**:* First, we examine the performance in terms of *decoding rate*, i.e., the percentage of transmitted bits that can be correctly decoded by the receiver. Figure 13 exhibits the achieved decoding rate as a function of β and symbol length for the four music traces. As expected, the decoding rate can be improved by: (1) using a larger β, and (2) by applying the frequency masking theory. Intuitively, a larger β will sacrifice the encoding rate, but allows a data bit to be transmitted at high energy level for a longer time. Second, given the symbol length and β, HiddenCode-Mask achieves higher decoding rate, as it is able to transmit the data bit at a higher signal energy. As shown in Figure 13, with 100ms symbol intervals, HiddenCode-Mask is able to maintain a decoding rate above 98% for $\beta \geq 0.3$, whereas, β needs to be greater than 0.4 with 50ms symbol interval. In case of HiddenCode-NonMask, it requires $\beta \geq 0.4$, and $\beta \geq 0.5$, for 100ms and 50ms symbol interval, respectively.

Second, we examine the *effective data rate*, R_{data}, i.e., the number of bits that can be transmitted within one second. R_{data} is obtained by: $R_{data} = \frac{1s}{t_{symbol}} \times R_{encode}$, in which t_{symbol} is the symbol interval length, and R_{encode} is the encoding chance (the encoding rates of the four examined musics are given in Table II). Given a minimum decoding rate requirement of 98%, the achieved effective data rates for the four examined musics with different symbol lengths and β are shown in Table III. For the four examined musics, the proposed

TABLE III: The achieved data rate for the four musics given different symbol lengths and β. Notation '-' indicates the decoding rate is below the requirement of 98% in that case.

Trace	Symbol Length	β				
		0.1	0.2	0.3	0.4	0.5
Music 1	50ms	-	-	-	-	**7.6**
	100ms	-	-	4.6	4.2	3.8
Music 2	50ms	-	-	-	-	**8.4**
	100ms	-	-	4.7	4.5	4.5
Music 3	50ms	-	-	-	-	**14.2**
	100ms	-	-	7.8	7.6	7.6
Music 4	50ms	-	-	-	**10.4**	9.8
	100ms	-	-	6.5	5.3	4.9

system is able to achieve an effective data rate from 7.6 to 14.2 bps by using the current setting. Now, given the length of Data Sub-Block for the four examined music traces equals to 10.4, 10, 18, and 9.9 seconds, respectively, we can calculate the number of bits that can be effectively transmitted within the given transmission block. Based on the effective data rate reported in Table III, the proposed HiddenCode-Mask is able to transmit 79, 84, 255, and 103 bits of data within a single transmission block. This means that, HiddenCode is able to transmit a coupon with 9 to 31 character length (e.g., Walmart coupon codes have eight characters [33]).

C. Subjective Perception Assessment

In the following, we conduct a user study to investigate the perception of HiddenCode and compare it with Vibration-OOK (VOOK) [4].

1) Settings and Instructions: In terms of HiddenCode, we consider both **HiddenCode-Mask** and **HiddenCode-NonMask**. The symbol interval is configured to 100ms with β equals to 0.5. In case of VOOK, we follow the settings described in [4]: the bit '1' is represented by transmitting a pure tone at the resonance frequency of the VEH (i.e., 205Hz), while no signal is transmitted for bit '0'. Moreover, we adjust the signal amplitude of bits '1' at four different levels: 0.1, 0.2, 0.3, and 0.4 (denoted by **VOOK-1**, **VOOK-2**, **VOOK-3**, and **VOOK-4**, respectively). We also use a bandstop filter to filter out inband signal in the original audio (i.e., from 180 to 220Hz) to eliminate its interference. Finally, filtered signal is combined with the generated binary bits signal. The symbol interval we used is 100ms, and the transmission time for '1' is configured to 10ms which is the minimum required length [4].

We generate six sound traces by applying the above six modulation settings. Then, together with the original audio trace, we have seven sound traces to be examined by the subjects for each of the four music traces. Before the user study, the audio traces are randomly numbered from 1 to 7, and a blackbox testing is performed (i.e., the corresponding modulation setting is unknown by the subjects). We encourage readers to listen and assess the sound traces by themselves. All audio files used in this paper are accessible online [34].

In total, 30 volunteers have participated in our study. They are diverse in gender (18 females and 12 males), age (range for 22 to 36), and educational background (i.e., with and without knowledge in computer science). The test is conducted in a

TABLE IV: The averaged MOS scores with the corresponding decoding rate shown in the bracket alongside.

Method	Music 1	Music 2	Music 3	Music 4	Average
VOOK-1	4.73 (20%)	3.46 (13%)	4.73 (10%)	4.53 (5%)	4.36 (12%)
VOOK-2	3.60 (83%)	1.60 (86%)	3.47 (87%)	2.47 (81%)	2.79 (84%)
VOOK-3	2.07 (100%)	1.53 (96%)	2.20 (96%)	2.33 (100%)	2.03 (98%)
VOOK-4	1.80 (100%)	1.46 (100%)	2.00 (100%)	1.87 (100%)	1.78 (100%)
HiddenCode-NonMask	4.60 (100%)	4.73 (100%)	4.33 (100%)	4.67 (100%)	4.58 (100%)
HiddenCode-Mask	4.53 (100%)	4.40 (100%)	4.53 (100%)	4.73 (100%)	4.55 (100%)
Original Audio	4.73	4.53	4.80	4.87	4.73

quiet office room, and subjects are asked to wear the Bose QuietComfort 35 noise cancellation headphones to ensure high audio quality and minimize surrounding noise. The headphone is connected to a laptop where all the audio traces will be played. We use the Mean Opinion Score (MOS) [35] as the metric to evaluate the subject perception. After listening each of the audio traces, subjects are asked to give a score from 1 to 5 [2], which indicates "Annoying", "Slightly Annoying", "Not Annoying", "Almost Unnoticeable", and "Completely Unobtrusive", respectively. All the 30 subjects performed the test independently, and they listened all the audio traces within one session which lasts less than 10 minutes.

2) Results Analysis: The averaged MOS scores for the four musics with different modulation settings are shown in Table IV. The corresponding decoding rates are shown alongside with the MOS scores. In terms of VOOK, the MOS score decreases from VOOK-1 to VOOK-4 with the increase of the signal amplitude. The MOS scores for VOOK-2, VOOK-3, and VOOK-4, are all below 3. This means that the embedded signal are annoying to the subjects in those settings. Although, the average score for VOOK-1 is above 4 and made the embedded signal 'almost unnoticeable' to the subjects, the averaged decoding rate for VOOK-1 is only 12% across the four musics. This makes VOOK-1 completely useless for data communication. In contrast, for all the examined musics, the MOS scores for both HiddenCode-NonMask and HiddenCode-Mask are all above 4, while maintain 100% decoding rate. This suggests that HiddenCode is able to provide VEH-based data communication without user perception.

VII. Discussion

In this work we used an off-the-shelf VEH consisting of a *single* cantilever made from *bulk* piezoelectric material. Although it allowed us to demonstrate the potential of the proposed musical masking techniques for VEH-based audio communications, the achieved data rates and the distance between the speaker and the VEH receiver were limited. In this section, we discuss two interesting trends in VEH research that may alleviate these limitations in the future.

A. Improving Data Rate with VEH Array

Researchers have recently demonstrated that it is possible to micromachine *multiple* piezoelectric cantilevers in a single substrate to produce an array of multiple individual energy harvesters in a single micro-electro-mechanical system (MEMS) [36], [37], [38], [39]. Although such VEH arrays are purely motivated by the need to produce more energy from the harvesting device, it opens up a fascinating new opportunity for VEH-based audio communication. A VEH array can be paralleled to some extent with the so called *multiple antenna systems*, such as multiple input and multiple output (MIMO), used in wireless communications to increase the received signal quality. For example, a VEH array configured with different resonance frequencies for different cantilevers, could be potentially used for multiple channel communication, which captures the encoded acoustic signal transmitted at different frequencies, thus providing at least linear improvement in the effective data rate.

B. Nanotechnology-based ultra-sensitive VEH

Existing bulk materials have limited sensitivity to vibrations. They can only react to vibrations with amplitudes higher than a given threshold. As such, our capability to sense acoustic signals using VEH is limited by the threshold properties of the bulk material. With advancements in nanotechnology, it is becoming possible to manufacture materials at nanoscale and engineer molecular structures that can achieve far greater sensitivity than bulk materials. Piezoelectric ZnO nanowire [40], [41] is one such recent wonders of nanotechnology that offers extreme sensitivity to vibrations of ultra-small magnitude for mechanical energy harvesting purposes. These technologies are still in research laboratories, but once they become available to the mass market, we will have the opportunity to detect ultra-small acoustic vibrations readily from the energy harvesters. Such ultra-sensitive signals may help us to significantly improve the transmission distance.

VIII. Conclusion

We have shown that musical masking of VEH-based audio communication has the unique characteristic of not being able to encode information in all symbols within a data packet. Hiding VEH communication within music therefore requires a solution for the decoder to learn the symbol patterns used by the encoder. We have proposed, implemented, and evaluated an encoding technique, which capitalises sound repetitions used in popular songs to capture the listener's attention. The encoder transmits the first occurrence of the repeated sound unencoded allowing the decoder to retrieve the data symbol patterns used by the encoder in subsequent repetitions. As a result, data can be effectively encoded and decoded from non-consecutive symbols in a packet, thus overcoming the challenges involved in musical masking of VEH communications. The effectiveness of hiding the communications sound from the humans has been verified by subjective evaluations. The outcome will contribute to more pervasive deployments of VEH-based audio communications.

REFERENCES

[1] C. V. Lopes and P. M. Aguiar, "Acoustic modems for ubiquitous computing," *IEEE Pervasive Computing*, vol. 2, no. 3, pp. 62–71, 2003.

[2] Q. Wang, K. Ren, M. Zhou, T. Lei, D. Koutsonikolas, and L. Su, "Messages behind the sound: real-time hidden acoustic signal capture with smartphones," in *Proceedings of MobiCom*. ACM, 2016, pp. 29–41.

[3] Q. Lin, L. Yang, and Y. Liu, "Tagscreen: Synchronizing social televisions through hidden sound markers," in *Proceedings of INFOCOM*. IEEE, 2017.

[4] G. Lan, W. Xu, S. Khalifa, M. Hassan, and W. Hu, "Veh-com: Demodulating vibration energy harvesting for short range communication," in *Proceedings of PerCom*. IEEE, 2017.

[5] "AMPY," http://www.getampy.com/ampy-move.html//.

[6] "SOLEPOWER," http://www.solepowertech.com/.

[7] "KINERGIZER," http://kinergizer.com/.

[8] "Kinetic," http://www.seiko-cleanenergy.com/watches/kinetic-1.html.

[9] S. Sudevalayam and P. Kulkarni, "Energy harvesting sensor nodes: Survey and implications," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 3, pp. 443–461, 2011.

[10] R. J. Vullers, R. Van Schaijk, H. J. Visser, J. Penders, and C. Van Hoof, "Energy harvesting for autonomous wireless sensor networks," *IEEE Solid-State Circuits Magazine*, vol. 2, no. 2, pp. 29–38, 2010.

[11] F. Li, T. Xiang, Z. Chi, J. Luo, L. Tang, L. Zhao, and Y. Yang, "Powering indoor sensing with airflows: a trinity of energy harvesting, synchronous duty-cycling, and sensing," in *Proceedings of SenSys*. ACM, 2013, p. 73.

[12] S. Khalifa, M. Hassan, and A. Seneviratne, "Pervasive self-powered human activity recognition without the accelerometer," in *Proceedings of PerCom*. IEEE, 2015, pp. 79–86.

[13] W. Xu, G. Lan, Q. Lin, S. Khalifa, N. Bergmann, M. Hassan, and W. Hu, "Keh-gait: Towards a mobile healthcare user authentication system by kinetic energy harvesting," in *Proceedings of NDSS*, 2017.

[14] S. Khalifa, G. Lan, M. Hassan, A. Seneviratne, and S. K. Das, "Harke: Human activity recognition from kinetic energy harvesting data in wearable devices," *IEEE Transactions on Mobile Computing (to appear)*, 2017, DOI:10.1109/TMC.2017.2761744.

[15] J. Twiefel and H. Westermann, "Survey on broadband techniques for vibration energy harvesting," *Journal of Intelligent Material Systems and Structures*, vol. 24, no. 11, pp. 1291–1302, 2013.

[16] C. Harrison, D. Tan, and D. Morris, "Skinput: appropriating the body as an input surface," in *Proceedings of CHI*. ACM, 2010, pp. 453–462.

[17] A. Madhavapeddy, R. Sharp, D. Scott, and A. Tse, "Audio networking: the forgotten wireless technology," *IEEE Pervasive Computing*, vol. 4, no. 3, pp. 55–60, 2005.

[18] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.

[19] W. A. Yost, "Fundamentals of hearing: an introduction," 2001.

[20] Y. Lin, W. H. Abdulla *et al.*, *Audio Watermark*. Springer, 2015, vol. 146.

[21] H. Lee, T. H. Kim, J. W. Choi, and S. Choi, "Chirp signal-based aerial acoustic communication for smart devices," in *Proceedings of INFOCOM*. IEEE, 2015, pp. 2407–2415.

[22] S. Ka, T. H. Kim, J. Y. Ha, S. H. Lim, S. C. Shin, J. W. Choi, C. Kwak, and S. Choi, "Near-ultrasound communication for tv's 2nd screen services," in *Proceedings of MobiCom*. ACM, 2016, pp. 42–54.

[23] L. Boney, A. H. Tewfik, and K. N. Hamdy, "Digital watermarks for audio signals," in *Multimedia Computing and Systems, 1996., Proceedings of the Third IEEE International Conference on*. IEEE, 1996, pp. 473–480.

[24] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on selected areas in communications*, vol. 6, no. 2, pp. 314–323, 1988.

[25] "Billboard top 200," http://www.billboard.com/charts/billboard-200, accessed: 2017-08-10.

[26] "Mide," http://www.mide.com/, accessed: 2017-09-19.

[27] D. J. Hargreaves, "The effects of repetition on liking for music," *Journal of research in Music Education*, vol. 32, no. 1, pp. 35–47, 1984.

[28] E. H. Margulis, *On repeat: How music plays the mind*. Oxford University Press, 2014.

[29] P. Kivy, *The fine art of repetition: Essays in the philosophy of music*. Cambridge University Press, 1993.

[30] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[31] G. Lan, D. Ma, W. Xu, M. Hassan, and W. Hu, "Capsense: Capacitor-based activity sensing for kinetic energy harvesting powered wearable devices," in *Proceedings of MobiQuitous*, 2017.

[32] "Sound pressue level table," http://www.sengpielaudio.com/, accessed: 2017-09-14.

[33] "Walmart coupon code," https://slickdeals.net/coupons/walmart-to-go/, accessed: 2017-09-25.

[34] Audio files used in this paper. [Online]. Available: https://www.dropbox.com/sh/75vd2b642nwjreq/AAAocobqBwU-WlPCJhhe-B-9a?dl=0

[35] A. W. Rix, "Perceptual speech quality assessment-a review," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 3. IEEE, 2004, pp. iii–1056.

[36] H. Yu, J. Zhou, L. Deng, and Z. Wen, "A Vibration-Based MEMS Piezoelectric Energy Harvester and Power Conditioning Circuit," *Sensors*, vol. 14, pp. 3323–3341, 2014.

[37] S. Roundy, E. S. Leland, J. Baker, E. Carleton, E. Reilly, E. Lai, B. Otis, J. M. Rabaey, P. K. Wright, and V. Sundararajan, "Improving power output for vibration-based energy scavengers," *IEEE Pervasive computing*, vol. 4, no. 1, pp. 28–36, 2005.

[38] J.-Q. Liu, H.-B. Fang, Z.-Y. Xu, X.-H. Mao, X.-C. Shen, D. Chen, H. Liao, and B.-C. Cai, "A mems-based piezoelectric power generator array for vibration energy harvesting," *Microelectronics Journal*, vol. 39, no. 5, pp. 802–806, 2008.

[39] H. Xue, Y. Hu, and Q.-M. Wang, "Broadband piezoelectric energy harvesting devices using multiple bimorphs with different operating frequencies," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 55, no. 9, 2008.

[40] Z. L. Wang and J. Song, "Piezoelectric nanogenerators based on zinc oxide nanowire arrays," *Science*, vol. 312, no. 5771, pp. 242–246, 2006.

[41] S. Xu, Y. Qin, C. Xu, Y. Wei, R. Yang, and Z. L. Wang, "Self-powered nanowire devices," *Nature nanotechnology*, vol. 5, no. 5, pp. 366–373, 2010.