Listening through a Vibration Motor

Nirupam Roy, Romit Roy Choudhury
University of Illinois at Urbana-Champaign

ABSTRACT

This paper demonstrates the feasibility of using the vibration motor in mobile devices as a sound sensor, almost like a microphone. We show that the vibrating mass inside the motor – designed to oscillate to changing magnetic fields – also responds to air vibrations from nearby sounds. With appropriate processing, the responses become intelligible, to the extent that humans can understand the vibra-motor recorded words with greater than 80% average accuracy. Even off-the-shelf speech recognition softwares are able to decode at 60% accuracy, without any training or machine learning. While these findings are not fundamentally surprising (given that any vibrating object should respond to air vibrations), the fidelity to which this is possible has been somewhat unexpected. We present our overall techniques and results through a system called *VibraPhone*, and discuss implications to both sensing and security.

1. INTRODUCTION

Vibration motors, also called "vibra-motors", are small actuators embedded in all types of phones and wearables. These actuators have been classically used to provide tactile alerts to human users. This paper identifies the possibility of using vibra-motors as a sound sensor, based on the observation that the same movable mass that causes the pulsation, should also respond to changes in air pressure. Even though the vibra-motor is likely to be far less sensitive compared to the (much lighter) diaphragm of an actual microphone, the question we ask is: *to what fidelity can the sound be reproduced?*

Even modest reproduction could prompt new applications and threats. On one hand, wearable devices like fitbits, that otherwise do not have a microphone, could now respond to voice commands. Further, in devices that already have microphones, perhaps better SNR could be achieved by combining the uncorrelated (noise) properties of the vibra-motor and microphone. On the other hand, leaking sound through vibra-motors opens new side channels – a malware that has default access to a phone's vibra-motor may now be able to eavesdrop into every phone conversation. Toys that have vibra-motors embedded could potentially listen into the ambience. This paper is an investigation into the vibra-motor's efficacy as a sound sensor, speech in particular.

Our work follows a recent line of work in which motion sensors in smartphones have been shown to detect sound. Authors of Gyrophone [30] first demonstrated the feasibility of detecting sound signals from the rotational motions of smartphone gyroscopes. A recent work [47] reported how accelerometers may also be able to detect sound, in fact, classify spoken keywords such as "OK Google" or "Hello Siri". Authors rightly identified the applicability to continuous sound sensing – the energy-efficient accelerometer could always stay active, and turn on the energy-hungry microphone only upon detecting a keyword. While certainly useful, we observe that these systems run pattern recognition algorithms on the features of the signals. The vocabulary is naturally limited to less than 3 keywords, trained by a specific speaker. VibraPhone is attempting a different problem altogether – instead of learning a motion signature, it attempts to reconstruct the inherent speech content from the low bandwidth, highly distorted output of the vibra-motor. Hence, there are no vocabulary restrictions, and the output of VibraPhone should be decodable by speech-to-text softwares.

As a first step towards converting a vibra-motor into a sound sensor, VibraPhone exploits the notion of *reverse electromotive force* (back-EMF) in electronic circuits. Briefly, the A/C current in the vibra-motor creates a changing magnetic field around a coil, which in turn causes the vibra-motor mass to vibrate. However, when an external force impinges on the same mass – say due to the pressure of ambient sound – it causes additional motion, translating into a current in the opposite direction. This current, called back-EMF, can be detected through an ADC after sufficient amplification. Of course, the signal extracted from the back-EMF is noisy and at a lower bandwidth than human speech. However, given that human speech obeys an "acoustic grammar", we find an opportunity to recover the spoken words even from the back-EMF's signal traces. VibraPhone focuses on exactly this problem, and develops a sequence of techniques, including *spectral subtraction, energy localization, formant extrapolation, and harmonic reconstruction*, to ultimately distill out legible speech.

Our experimentation platform is both a Samsung smartphone and a custom circuit that uses vibra-motor chips pur-

MobiSys'16, June 25 - 30, 2016, Singapore, Singapore

ISBN 978-1-4503-4269-8/16/06...$15.00

DOI: http://dx.doi.org/10.1145/2906388.2906415

chased online (these chips are exactly the ones used in today's phones and wearables). We characterize the extent of signal reconstruction as a function of the loudness of the sound source. Performance metrics are defined by the accuracy with which the reconstructed signals are intelligible to humans and to (open-source) automatic speech recognition softwares. We use the smartphone microphone as an upper bound, and for fairness, record the speech at the same sound pressure level (SPL) [24, 4, 42] across all the devices. We experiment across a range of scenarios within our university building, and observe that results are robust/useful when the speaker is less than 2 meters from the vibra-motor.

Finally, we emphasize that smartphone vibra-motors cannot be used as microphones today, primarily because the actuator is simply not connected to an ADC. To this end, launching side-channel attacks is not immediate. However, as discussed later, we find that enabling the listening capability requires almost trivial rewiring (just soldering at 4 clearly visible junctions). This paper sidesteps these immediacy questions and concentrates on the core nature of the information leakage. At the least, we hope this work will draw attention to the permission policies on vibra-motors, which today are open to all apps by default. We have made various audio demos of VibraPhone available on our website [5] – we request the readers to listen to them to better experience the audio effects and reconstructions. In closing, the main contributions in this paper may be summarized as:

- *Recognizing that ambient sound manifests itself as back-EMF inside vibra-motor chips.* This leads to an actuator becoming a sound sensor with minimal changes to the current mobile device hardware.
- *Designing techniques that exploit constraints and structure of human speech to decode words from a noisy, low bandwidth signal.* Building the system on a smartphone and custom hardware platform, and demonstrating decoding accuracy of up to 88% when a male user is speaking in normal voice near his phone.

The rest of the paper expands on these contributions. We begin with a brief introduction to vibra-motors and our hardware platform.

2. UNDERSTANDING VIBRA-MOTORS

A vibra-motor is an electro-mechanical device that moves a magnetic mass rhythmically around a neutral position to generate vibrations [36]. While there are various kinds of vibra-motors, a popular one is called *Linear Resonant Actuators (LRA)* shown in Figure 1. With LRA, vibration is generated by the linear movement of the magnetic mass suspended near a coil, called the "voice coil". Upon applying AC current to the motor, the coil also behaves like a magnet (due to the generated electromagnetic field) and causes the mass to be attracted or repelled, depending on the direction of the current. This generates vibration at the same frequency as the input AC signal, while the amplitude of vibration is dictated by the signal's peak-to-peak voltage. Thus LRAs offer control on both the magnitude and frequency of vibration. Most smartphones today use LRA based vibra-motors.

2.1 Sound Sensing through back-EMF

Back-EMF is an electro-magnetic effect observed in magnet-based motors when relative motion occurs between the current carrying armature/coil and the magnetic mass's own

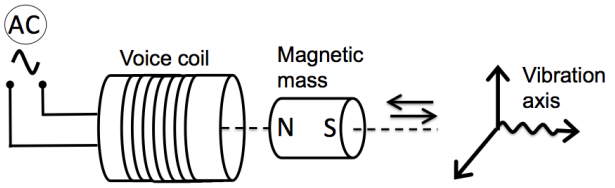

Figure 1: Basic sketch of an LRA vibra-motor.

field. According to Faraday's law of electromagnetic induction [16], this changing magnetic flux induces an electromotive force in the coil. Lenz's law [41] says this electromotive force acts in the reverse direction of the driving voltage, called *back-EMF of the motor*. As the rate of change of the magnetic flux is proportional to the speed of the magnetic mass, the back-EMF serves as an indicator of the extraneous vibration experienced by the mass.

Since sound is a source of external vibration, the movable mass in the vibra-motor is expected to exhibit a (subtle) response to it. Our experiments show that, when the vibra-motor is connected to an ADC, the back-EMF generated by the ambient sound can be recorded. This is possible even when the vibra-motor is passive (i.e., not pulsating to produce tactile alerts). We call this ADC output *vibra-signal* to distinguish it from the microphone signal that we will later use as a baseline for comparison. We now describe our platform to record and process the vibra-signal.

2.2 Experiment Platform

Custom hardware setup: Today's smartphones offer limited exposure/API to vibra-motor capabilities and other hardware components (e.g., amplifiers). To bypass these restrictions, we have designed a custom hardware setup using off-the-shelf LRA vibra-motor chips connected to our own ADC and amplifier. Figure 2 shows our setup – we mount this vibration motor adjacent to a standard microphone that serves as a comparative baseline. The vibra-signal is amplified and sampled at 16KHz. Test sounds include live speech from humans at varying distances, as well as sound playbacks through speakers at varying loudness levels.

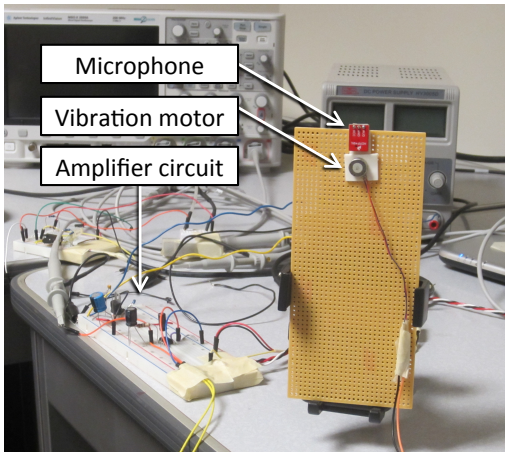

Figure 2: The custom hardware setup with collocated vibration motor and microphone.

Smartphones: While the custom hardware offers better programmability, we also use a smartphone setup to understand the possibilities with today's systems. Figure 3 shows

our prototype – terminals of the built-in vibra-motor of a Samsung Galaxy S-III smartphone is connected to the audio line-in input port with a simple wire. The rewiring is trivial – for someone familiar with the process, it can be completed in less than 10 minutes. Once rewired, we collect the samples of the vibra-signal from the output channels of the earphone jack, using our custom Android application.

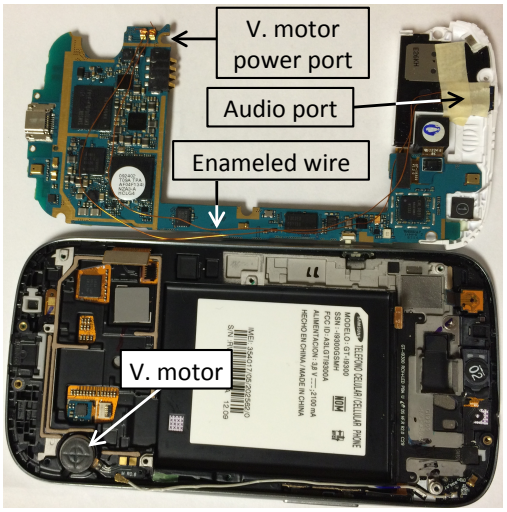

Figure 3: The smartphone setup with a simple wire connected between the vibra-motor's output to the audio line-in port.

Electromagnetic Coupling: We conduct a microbenchmark test to verify that the vibration motor signal is not influenced by the Electromagnetic coupling from the nearby microphone or speakers in our test setup. We remove the speakers and microphones from the test environment and directly record human speech with a vibration motor (find sample clips at project website [5]). Later we compare them with the recordings of the standard test setup to find no noticeable difference in signal quality.

3. SOUNDS AND HUMAN SPEECH

This section is a high level introduction to speech production in humans, followed by a discussion on the structure of speech signals.

3.1 Human Speech Production

Human speech can be viewed as periodic air waves produced by the lungs, modulated through a sequence of steps in the throat, nose, and mouth. More specifically, the air from the lungs first passes through the *vocal cords* – a pair of membranous tissue – that constricts or dilates to block or allow the air flow (Figure 4). When the vocal cords are constricted, the vibrations induced in the air-flow are called *voiced* signals. The voiced signals generate high energy pulses – in the frequency domain, the signal contains a fundamental frequency and its harmonics. All vowels and some consonants like "b" and "g" are sourced in voiced signals.

On the other hand, when the vocal cords dilate and allow the air to flow through without heavy vibrations, the outcome is called *unvoiced* signals. This generates sounds similar to noise, and is the origin of certain consonants, such as "s", "f", "p", "k", "t". Both voiced and unvoiced signals then pass through a flap of tissue, called *glottis*, which further pulsates to add power to the signal as well as distinctiveness to an individual's voice. These *glottal pulses* travel further and are finally modulated by the oral/nasal cavities to produce fine-tuned speech [11]. The overall speech production process is often modeled as a "source-filter" in literature, essentially implying that the human trachea/mouth applies a series of filters to the source sound signal. This source-filter model will later prove useful, when VibraPhone attempts to reconstruct the original speech signal.

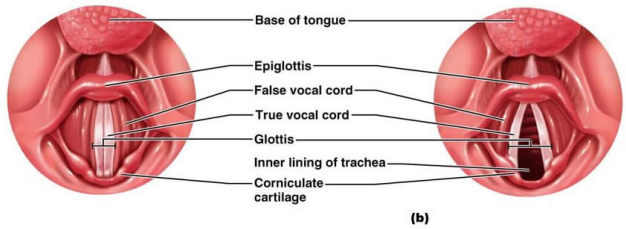

Figure 4: The vocal cords constricted in (a) and dilated in (b), creating *voiced* and *unvoiced* air vibrations, that are then shaped by the glottis and epiglottis.

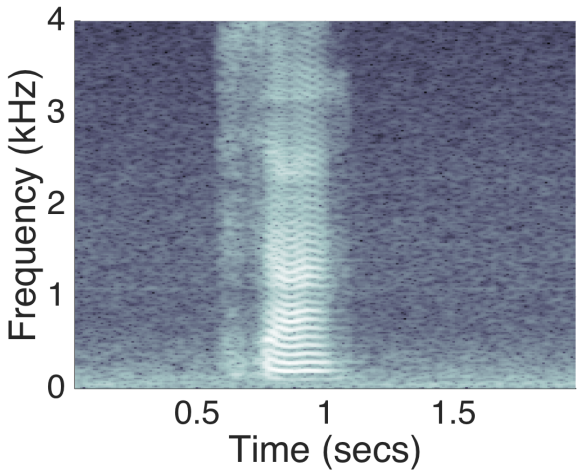

Figure 5: The spectrogram of the spoken consonant 's' followed by the vowel 'a' recorded with microphone.

3.2 Structure in Speech Signals

While the above discussions present a biological/linguistics point of view, we now discuss how they relate to the recorded speech signals and their structures. Figure 5 shows the spectrogram when a human user pronounces the alphabets "sa" – the signal was recorded through a smartphone microphone (not a vibra-motor)[1]. Although a toy case, the spectrogram captures the key building blocks of speech structure. We make a few observations that will underpin the challenges and the designs in the rest of the paper.

- The first visible signal (between 0.6 and 0.75 seconds) corresponds to the *unvoiced* component, the consonant "s". This signal is similar to noise with energy spread out rather uniformly across the frequency band. The energy content in this signal is low to moderate.

[1]The Y axis shows up to 4KHz, since normal human conversation in non-tonal languages like English is dominantly confined to this band.

- The second visible signal corresponds to the vowel "a" and is an example of the *voiced* component. The signal shows a low fundamental frequency and many harmonics all the way to 4KHz. Fundamental frequencies are around 85–180Hz for males and 165–255Hz for females [43]. The energy content of this signal is far stronger than the *unvoiced* counterpart.
- Within the *voiced* signal, the energy content is higher in the lower frequencies. These strong low frequency components determine the intelligibility of the spoken phonemes (i.e. the perceptually distinct units of sound [44]), and are referred to as *formants* [28]. The first two formants (say, $F1$, $F2$) remain between 300–2500Hz and completely forms the sound of the vowels, while some consonants have another significant formant, $F3$, at a higher frequency. Figure 6 shows examples of 2 vowel formants – "i" and "a" – recorded by the microphone.

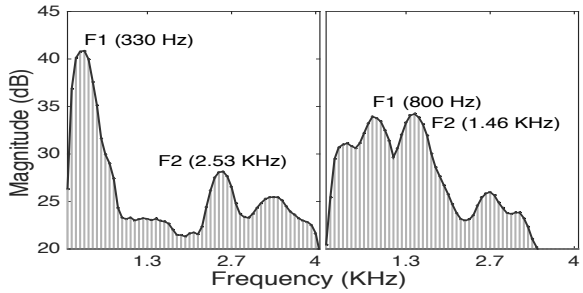

Figure 6: The locations of the first two formants (F1 and F2) for (a) the vowel sound *'i'* and (b) the vowel sound *'a'*, both recorded with microphone.

In extracting human speech from the vibra-motor's back-EMF signal, VibraPhone will need to identify, construct, and bolster these formants through signal processing.

4. CHALLENGES

Figure 7(a,b) compares the spectrogram of the microphone and the vibra-motor for the same spoken phoneme, "sa". Figure 7(c,d) shows the same comparison for a full word, namely, "entertainment". The reader is encouraged to listen to these sound clips at our project website [5]. Evidently, the vibra-motor's response is weak and incomplete, and on careful analysis, exhibits various kinds of distortions even where the signal is apparently strong. The goal in this paper is to reconstruct, to the extent possible, the left columns of Figure 7 from the right columns. We face 4 key challenges discussed next.

(1) Over-Sensitivity at Resonance Frequency

All rigid objects tend to oscillate at a fixed natural frequency when struck by an external force. When the force is periodically repeated at a frequency close to the object's natural frequency, the object shows exaggerated amplitude of oscillation – called *resonance* [34]. Resonance is often an undesirable phenomenon, destabilizing the operation of an electro-mechanical device. Microphones, for example, carefully avoid resonance by designing its diaphragm at a specific material, tension, and stiffness – that way, the resonance frequencies lie outside the operating region [19, 10]. In some cases, additional hardware is embedded to damp the vibration at the resonant frequencies [10].

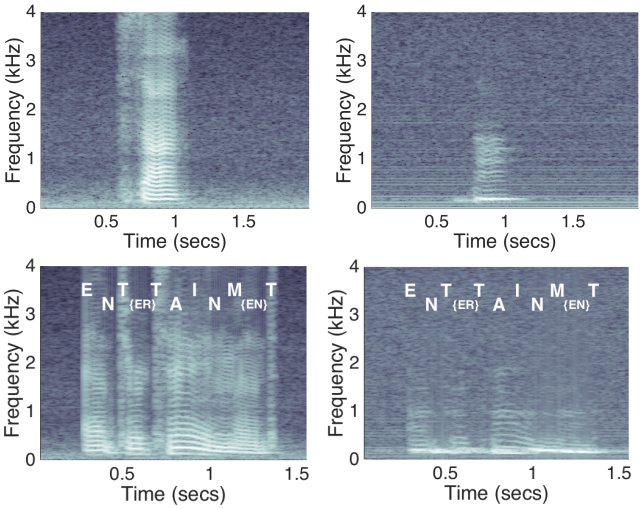

Figure 7: The spectrogram for "sa" as recorded by: (a) the microphone and (b) the vibra-motor. The spectrogram for the full word "entertainment" as recorded by: (c) the microphone and (d) the vibra-motor. The vibra-motor's response is weak and partially missing.

Unfortunately, vibra-motors used in today's smartphones exhibit sharp resonance between 216 to 232Hz, depending on the mounting structure. Some weak components of *speech formants* are often present in these bands – these components get amplified, appearing as a *pseudo-formant*. The *pseudo-formants* manifest as unexpected sounds within uttered words, affecting intelligibility. The impact is exacerbated when the fundamental frequency of the voiced signal is itself close to the resonant band – in such cases, the sound itself gets garbled. Figure 8 shows the effect of resonance when the vibration motor is sounded with different frequency tones in succession (called a *Sine Sweep* [13, 12]. Observe that for all tones in the *Sine Sweep*, the vibra-motor exhibited appreciable response in the resonance band. This is because the tones have some frequency tail around the 225Hz, and this always gets magnified. The microphone exhibits no such phenomenon. VibraPhone will certainly need to cope with resonance.

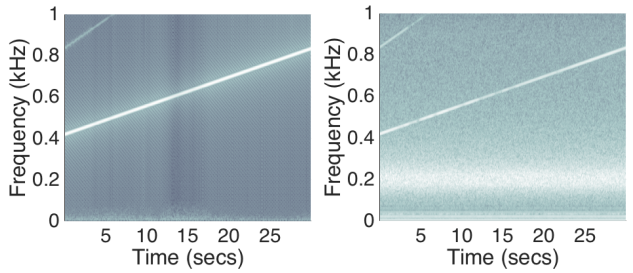

Figure 8: The spectrogram of (a) the microphone and (b) the vibra-motor, in response to a *Sine Sweep* (i.e., tones played at increasing narrow band frequencies). The vibration motor signal shows an over-sensitive resonance frequency band near 220Hz.

(2) High Frequency Deafness

The vibra-motor's effective diaphragm – the area amenable to the impinging sound – is around 10mm, almost 20x larger

than that of a typical MEMS microphone (0.5mm). This makes the vibration motor directional for the high frequency sounds, i.e., the high frequencies arriving from other directions are suppressed, somewhat like a directional antenna. Unfortunately, human voices contain lesser energy at frequencies higher than 2KHz, thereby making the vibra-motor even less effective in "picking up" these sounds. Some consonants and some vowels – such as "i" and "e" – have formants close to or higher than 2KHz, and are severely affected. Figure 9 compares the spectrogram when just the vowel "*a*" was spoken – evidently, the vibra-motor is almost "deaf" to higher frequencies.

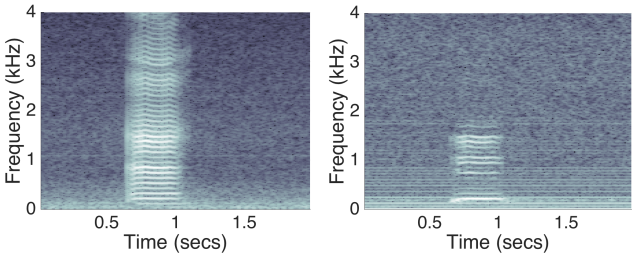

Figure 9: The spectrogram of the spoken vowel '*a*' recorded with (a) microphone and (b)vibration motor. The vibra-motor exhibits near-deafness for frequencies > 2KHz.

(3) Higher Energy Threshold

A microphone's sensitivity, i.e., the voltage produced for a given sound pressure level, heavily depends on the weight and stiffness of its diaphragm. The spring-mass arrangement of the vibra-motor is considerably more stiff, mainly due to the heavier mass and high spring constant. While this is desirable for a vibration actuator, it is unfavorable to sound sensing. Thus, using the actuator as a sensor yields low sensitivity in general, and particularly to certain kinds of low-energy consonants (like *f, s, v, z*), called *fricatives* [18]. The effect is visible in Figure 7 (a,b) – the *fricative* consonant "*s*" goes almost undetected with vibra-motors.

(4) Low Signal-to-Noise Ratio (SNR)

In any electrical circuit, *thermal noise* is an unavoidable phenomena arising from the Brownian motion of electrons in resistive components. Fortunately, the low 26 Ohm terminal resistance in vibra-motors leads to 10dB lower thermal noise than the reference MEMS microphone. However, due to low sensitivity, the strength of the vibra-signal is significantly lower, resulting in poor SNR across most of the spectrum. Figure 10 compares the SNR at different sound pressure levels – except around the resonance frequencies, the SNR of the vibra-signal is significantly less compared to the microphone.

Sound Pressure Level (SPL) is a metric to measure the effective pressure caused by sound waves with respect to a reference value, and is typically expressed in dBSPL [4]. This gives a standard estimate of the sound field at the receiver, irrespective of the location of the sound source.

5. SYSTEM DESIGN

Our system design is modeled as a *source-filter*, i.e., we treat the final output of the vibra-motor as a result of many filters applied serially to the original air-flow from the lungs. Figure 11 illustrates this view, suggesting that an ideal solution should perform two broad tasks: (1) "undo" the vibra-motor's distortions for signal components that have been detected, and (2) reconstruct the undetected signals by leveraging the predictable speech structure in conjunction with the slight "signal hints" picked up by the vibra-motor. VibraPhone realizes these tasks through two corresponding modules, namely, *signal pre-processing* and *partial speech synthesis*. We describe them next.

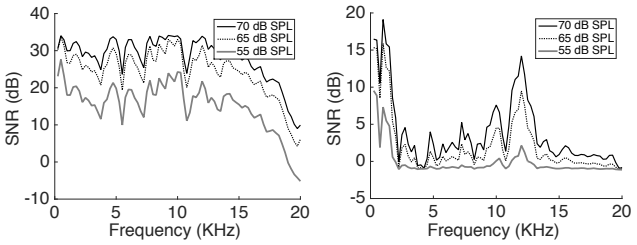

Figure 10: The SNR of (a) the microphone and (b) the vibra-motor at various frequencies for varying sound pressure levels (dB SPL). Note the unequal Y axis range.

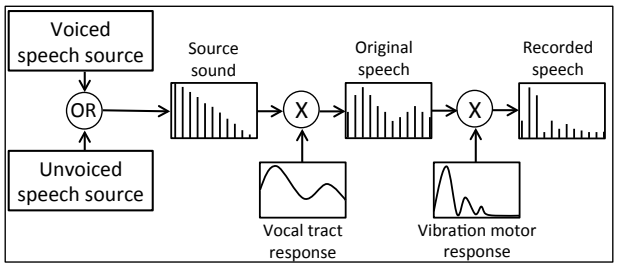

Figure 11: The source-filter model of the speech generation and recording.

5.1 Signal Pre-processing

All of our algorithms operate on the frequency domain representation of the signal. Therefore, we first convert the amplified signal to the time-frequency domain using the *Short Time Fourier Transform* (STFT), which basically computes the complex FFT coefficients from 100 millisecond segments (80% overlapped, Hanning windowed) of the input time signal. The result is a 2D matrix that we call time-frequency signal and illustrated in Figure 12 – each column is a time slice and each row is a positive frequency bin. We will refer to this matrix for various explanations.

Frequency Domain Equalization

When a microphone is subject to a *Sine Sweep* test, the frequency response is typically flat, meaning that the microphone responds almost uniformly to each frequency component. The vibra-motor's response, on the other hand, is considerably jagged, and thereby induces distortions into the arriving signal. Figure 13 shows a case where the vowel "u" is recorded by both the microphone and vibra-motor. The vibra-motor distortions on "u" are quite dramatic, altering the original formants at 266 and 600Hz to new formants at 300Hz and 1.06KHz. In fact, the altered formants bear resemblance to the vowel "aa" (as in "father"), and in reality, do sound like it. More generally, the vibra-motor's frequency

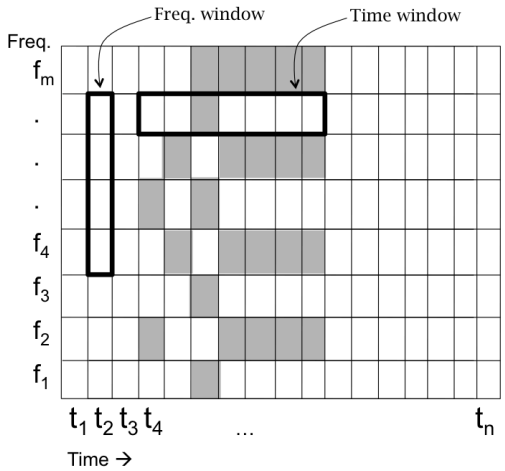

Figure 12: 2D time-frequency matrix

response exhibits this rough shape, thereby biasing all the vowels to the sound of "aa" or "o".

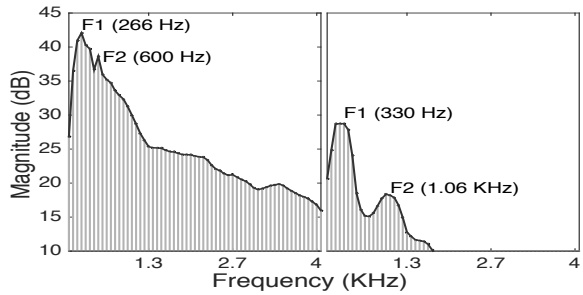

Figure 13: Formants of vowel *'u'* recorded through (a) microphone and (b) vibra-motor. The vibra-motor introduces a spurious formant near 1KHz.

Fortunately, the frequency response of the vibra-motor is only a function of the device and does not change with time (at least until there is wear and tear of the device). We tested this by computing the correlation of the *Sine Sweep* frequency response at various sound pressure levels – the correlation proved strong, except for a slight dip at the resonant frequencies due to the non-linearities. Knowing the frequency response, we apply an equalization technique, similar to channel equalization in communication. We estimate the inverse gain by computing the ratio of the coefficients from the microphone and the vibra-motor, and multiply the inverse gain with the frequency coefficients of the output signal.

Background Noise Removal

Deafness in vibra-motors implies that the motor's response to high frequency signals (i.e., > 2KHz) is indistinguishable from noise. If this noise exhibits a statistical structure, a family of *spectral subtraction* algorithms can be employed to improve SNR. However, two issues need attention. (1) The pure noise segments in the signal needs to be recognized, so that its statistical properties are modeled accurately. This means that noise segments must be discriminated from speech. (2) Within the speech segments, *voiced* and *unvoiced* segments must also be separated so that spectral subtraction is only applied on the *voiced* components. This is because *unvoiced* signals bear noise-like properties and spectral subtraction can be detrimental.

To reliably discriminate the presence of speech segments, we exploit the exaggerated behavior in the resonance frequency band. We consistently observed that speech brings out heavy resonance behavior in vibra-motors, while noise elicits a muted response. Thus, resonance proved to be an opportunity. Once speech is segregated from noise, the next step is to isolate the *voiced* components in speech. For this, we leverage its well-defined harmonic structure. Recall the 2D matrix in Figure 12. We consider a time window and slide it up/down to compute an autocorrelation coefficient across different frequencies. Due to the repetition of the harmonics, the autocorrelation spikes periodically, yielding robust detection accuracy. When autocorrelation does not detect such periodic spikes, they are deemed as the *unvoiced* segments.

The final task of spectral subtraction is performed on the *voiced* signal alone. For a given *voiced* signal (i.e., a set of columns in the matrix), the closest noise segments in time are selected – these noise segments are averaged over a modest time window. Put differently, for every frequency bin, the mean noise floor is computed, and then subtracted from the corresponding bin in the *voiced* signal. For zero mean Gaussian noise, this does not offer any benefit, however, the noise is often not zero mean. In such cases, the SNR improves and alleviates the deafness. Figure 14 shows the beneficial effect of spectral subtraction when "yes" is spoken.

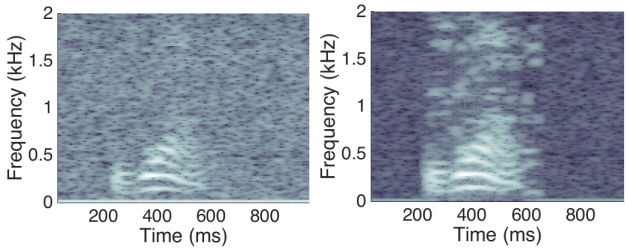

Figure 14: The spectrogram of the spoken word "yes" (a) before and (b) after the spectral subtraction.

Speech Energy Localization

Observe that noise removal described above brings the *mean noise* to zero, however, noise still exists and the SNR is still not adequate. In other words, deafness is still a problem. However, now that noise is zero mean and Gaussian, there is an opportunity to exploit its diversity to further suppress it. Even localizing the speech signal energy in the spectrogram would be valuable, even if the exact signal is not recovered in this step.

Our core idea is to average the signals from within a frequency window, and slide the frequency window all the way to 10KHz. Referring to the 2D matrix, we compute the average of W elements in each column (W being the window size), and slide the window vertically; the same operation is performed for each column. Each element is a complex frequency coefficient, containing both the signal and the noise. With sufficiently large W, the average converges to the average of the signal content in these elements since the (average) noise sum up to zero. Mathematically, if C_i denotes

the signal at frequency f_i, and $C_i = S_i + N_i$, where S_i is the speech signal and N_i the noise, then the averaged C_i^* is computed as:

$$C_i^* = \frac{1}{W} \sum_{f=i-\frac{W}{2}}^{i+\frac{W}{2}} C_i = \frac{1}{W} \sum_{f=i-\frac{W}{2}}^{i+\frac{W}{2}} S_i + \frac{1}{W} \sum_{f=i-\frac{W}{2}}^{i+\frac{W}{2}} N_i \quad (1)$$

Since the term $\sum N_i$ is zero mean Gaussian, it approaches zero for larger W, while the $\frac{1}{W}\sum S_i$ term is simple smoothing. For every frequency bin, we normalize the C_i^* values over a time window so that they range between $[0, 1]$. The result is a 3D contour map, where the locations of higher elevations, i.e., hills, indicate the presence of speech signals. We identify the dominant hills and *zero force* all areas outside them. This is because speech signals always exhibit a large time-frequency footprint, since human voice is not capable of producing sounds that are narrow in frequency and time. Figure 15 illustrates the effect of this scheme – the dominant hills are demarcated as the location of speech energy. Evidently, the improvement is conspicuous after this energy localization step.

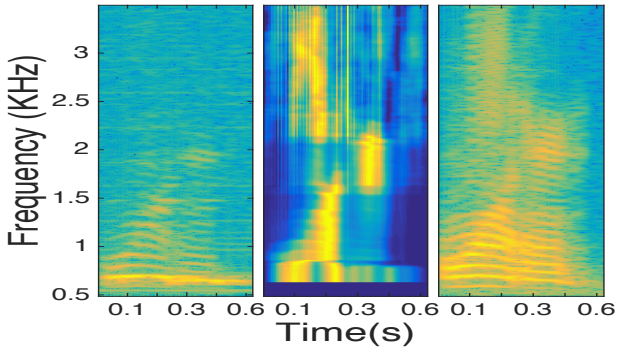

Figure 15: Readers are requested to view this figure in color: (a) Raw vibra-motor signal, (b) The output of the *speech energy localization* makes the signal energy visible through a heat-map like contour. (c) The corresponding microphone signal bearing good resemblance to the energy locations.

5.2 Partial Speech Synthesis

Once the vibra-motor output has been pre-processed, the structure of speech can now be leveraged for signal recovery – we describe our approach next.

Voice Source Expansion

After the localization step above, we know the location of speech energy (in time-frequency domain), but we do not know the speech signal. In attempting to recover this signal, we exploit the opportunity that the fundamental frequencies in speech actually manifest in higher frequency harmonics [35, 14]. Therefore, knowledge of the lower fundamental frequencies can be *expanded* to reconstruct the higher frequencies. Unfortunately, the actual fundamental frequency often gets distorted by the resonant bands.

As a workaround, we use the relatively high SNR signals in the range [250, 2000Hz] to synthesize the voice source signal at higher frequencies. Synthesis is essentially achieved through careful replication. Specifically, the algorithm copies the coefficient $C_{t,f}$, where t is the time segment and f is the frequency bin of the time-frequency signal, and adds it to $C_{t,kf}$ for all integer k, such that kf is less than the Nyquist frequency. Here integer k indicates the harmonic number for the frequency f. Intuitively, we are copying the harmonics from the reliable range, and replicating them at the higher frequencies. As shown in Figure 16, this only synthesizes the *voiced* components (recall the harmonics are only present in the *voiced* signals). For *unvoiced* signals, we blindly fill in the deaf frequencies with copies of the lower frequency signals.

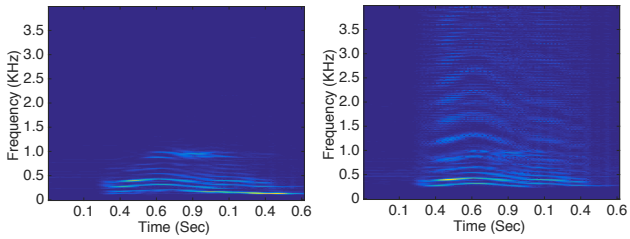

Figure 16: Result of source expansion for the *voiced* signal components. (a) Raw vibra-signal, and (b) after harmonic replication. Readers are requested to view this figure in color.

Speech Reconstruction

Recall that the mouth and nasal cavities finally modulate the air vibrations – this can be modeled as weights multiplied to the fundamental frequencies and their harmonics. While we do not know the values of these weights, the location of the energies – computed from the 3D contour hills – is indeed an estimate. We now utilize this location estimate as an *energy mask*. As a first step, we apply an exponential decay function along the frequency axis to model the low intensity of natural speech at the higher frequencies. Then the energy mask is multiplied with this modified signal, emulating an adaptive gain filter. As this also improves the SNR of the unvoiced section of the speech, we apply a deferred spectral subtraction method on these segments to further remove the background noise. Finally, we convert this resultant time-frequency signal to time domain using inverse short time Fourier transform (ISTFT). Figure 17 compares the output against the microphone and the raw vibra-motor signal.

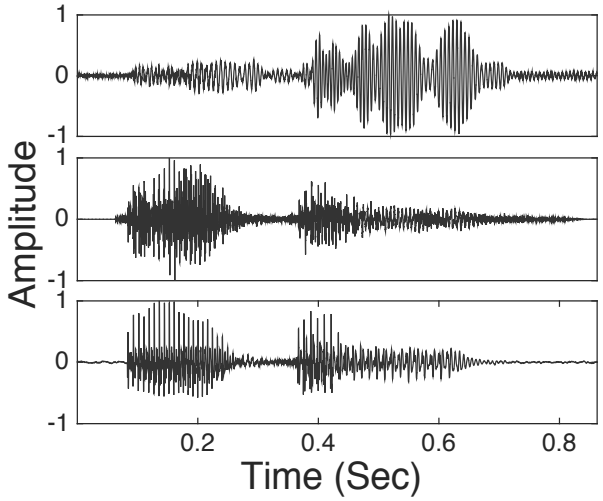

Figure 17: Word "often" as manifested in the (a) raw vibra-motor signal, (b) after VibraPhone's processing, and (c) microphone signal.

6. EVALUATION

Section 2.2 described the two experimentation platforms for our system, namely the custom hardware and the Samsung Galaxy smartphone. In both cases, we evaluate VibraPhone's speech intelligibility against the performance of the corresponding microphone. In the custom hardware, the microphone is positioned right next to the vibra-motor, while in the smartphone, their locations are modestly separated. We generate the speech signals using a text-to-speech (TTS) utility available in OS X 10.9, and play them at different volumes through a loudspeaker. The position/volume of the loudspeaker is adjusted such that the sound pressure levels at the vibra-motor and the microphone are equal. The accent and intonation of the TTS utility also does not affect the experiment since both VibraPhone and the microphone hear the same TTS speech. The content of the speech is drawn from Google's Trillion Word Corpus [3] – we pick 2000 most frequent words, which is prescribed as a good benchmark [33].

6.1 Methodology and Metrics

We perform automatic and manual speech recognition experiments as follows.

(1) Automatic Speech Recognition (ASR)

In ASR, a software programmatically converts the time domain speech signal to text. ASR tools typically have 3 distinct components: (a) an acoustic model, (b) a pronunciation dictionary, and (c) a language model. The *acoustic model* is a trained statistical model (e.g., HMM, Neural Networks, etc. [15, 20]) that maps segments of the input waveform to a sequence of phonemes. These phonemes are then looked up in the *pronunciation dictionary*, which lists the candidate words (along with their possible pronunciations) based on the matching phoneme sequence. Among these candidates, the most likely output is selected using a grammar or a *language model*.

Our ASR tools is the open-source *Sphinx4* (pre-alpha version) library published by CMU [1, 21]. The acoustic model is sensitive to the recording parameters, including the bandwidth and the features of the microphone. Such parameters do not apply to vibra-motors, so we used a generic acoustic model trained with standard microphone data. This is not ideal for VibraPhone, and hence, the reported results are perhaps a slight under-estimate of VibraPhone's capabilities.

(2) Manual Speech Recognition (MSR)

We recruited a group of 6 volunteers from our department building – 1 native English speaker, 1 Indian faculty with English as first language, 2 Indian PhD students, and 2 Chinese PhD students. We played the vibra-motor and microphone outputs to all the participants simultaneously and collected their responses. In some experiments, volunteers were asked to *guess the word or phrase from the playback*; in other experiments, the volunteers were given a list of phrases and asked to pick the most likely one, including the option of "none of the above". All human responses were accompanied by a subjective clarity score – every volunteer expressed how intelligible the word was, even when he/she could not guess with confidence. Finally, in some experiments, volunteers were asked to guess first, and then guess again based on a group discussion. Such discussions served as a "prior" for speech recognition, and often the group consensus was different from the first individual guess.

Metrics

Across all experiments, 9 hours of sound was recorded and a total of 20,000 words were tested with ASR at various sound pressure levels (measured in $dBSPL$). For MSR, a total of 300 words and 40 phrases were played, resulting in more than 2000 total human responses. We report "Accuracy" as the percentage of words/phrases that were correctly guessed, and show its variation across different loudness levels (measured in dBSPL). We report "Perceived Clarity" as a subjective score reported by individuals, even when they did not decode the word with confidence. Finally, we report "Precision", "Recall", and "Fallout" for experiments in which the users were asked to select from a list. Recall that *precision* intuitively refers to "*what fraction of your guesses were correct*", and *recall* intuitively means "*what fraction of the correct answers did you guess*". We now present the graphs, beginning with ASR.

6.2 Performance Results with ASR

Accuracy v/s Loudness

Custom Hardware: Figure 18(a) reports the accuracy with ASR as a function of the sound pressure level (dbSPL), a standard metric proportionally related to the loudness of the sound. VibraPhone's accuracy is around 88% at 80 dbSPL, which is equivalent to the sound pressure experienced by the smartphone's microphone during typical (against the ear) phone call. The microphone's accuracy is obviously better at 95%, while the raw vibra-motor signal performs poorly at 43% (almost half of VibraPhone). Importantly, the preprocessing and the synthesis gains are individually small, but since intelligibility is defined as binary metric here, the improvement jumps up when applied together.

Once the loudness decreases at 60 dbSPL – comparable to a normal conversation 1 meter away from the microphone [2] – VibraPhone's accuracy drops to $\approx$ 60%. At lower sound pressure level, the accuracy drops faster since the vibra-motor's sensitivity is inadequate for "picking up" the air vibrations. However, the accuracy can be improved with training the acoustic model with vibra-motors (recall that with ASR, the training is performed through microphones, which is unfavorable to VibraPhone).

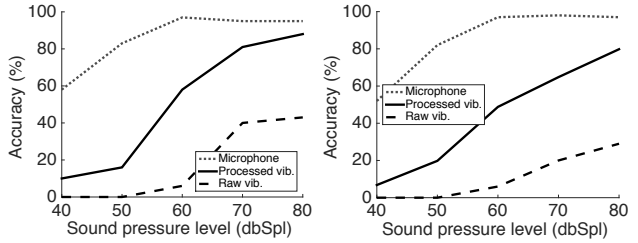

Figure 18: Automatic recognition accuracy as a function of loudness for (a) the custom hardware, (b) the Samsung smartphone.

Samsung Smartphone: Figure 18(b) plots the accuracy with ASR for the smartphone based platform. VibraPhone's

performance is weaker compared to the custom hardware setup, although the difference is marginal – ASR output is still at 80% at 80 dbSPL. Admittedly, we are not exactly sure of the reason for this difference – we conjecture that the smartphone signal processing pipeline may not be exactly tuned to the vibra-motor like we have done in the custom case.

Rank of the Words

The accuracy results above counts only perfect matches between ASR's output and the actual spoken word. In certain applications, a list of possible words may also be useful, particularly when the quality of the speech is poor. We record the list of all predictions from ASR for each spoken word, played at 50 dbSPL. Figure 19 plots the CDF of the rank of the correct word in this list. At this relatively softer 50 dbSPL experiment, only $\approx 20\%$ of the words are ranked at 1, implying exact match. In 41% of the cases, the words were within *top-5* of the list, and *top-10* presents a 58% accuracy.

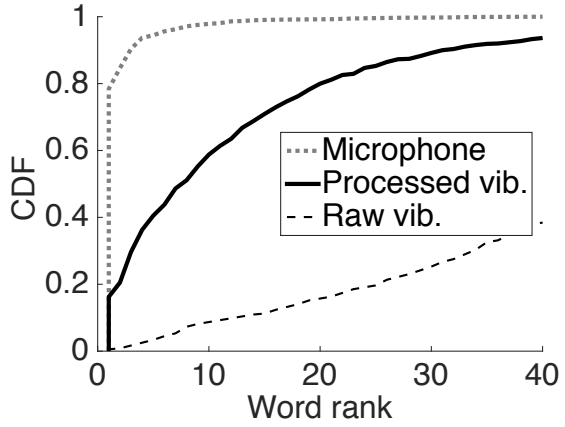

Figure 19: The CDF of word rank from ASR's prediction at 50 dbSPL for custom hardware.

Phoneme Similarity

The acoustic model we used with ASR is not ideal for VibraPhone – the impact is pronounced for distorted phonemes. Training ASR's acoustic model with the vibra-motor response is expected to offer improvements, but in the absence of that, we report a subjective overview of the entropy in different phonemes recorded by VibraPhone. In other words, we ask whether *autocorrelation* between the same phonemes is high and *cross correlation* across phonemes are low. We extract the STFT coefficients of the 100 phonemes (28 vowels and 72 consonants) from the International Phoneme Alphabet [6, 7] and use these coefficients as the features. We then calculate correlation coefficient of all pairs of phonemes in the list – Figure 20 presents the heat map. In case of raw vibra-signal in Figure 20(a), the (distorted) phonemes bear substantial similarity between each other, indicated by the multiple dark off-diagonal blocks. The two large darker squares in the figure represents the pulmonic (58 phonemes) and non-pulmonic (14 phonemes) consonant groups [27, 18]. However, with VibraPhone, Figure 20(b) shows substantial improvements. The autocorrelation is strong across the diagonal of the matrix, while the off-diagonal elements are much less correlated. This extends hope that a vibra-motor trained acoustic model could appreciably boost VibraPhone's performance.

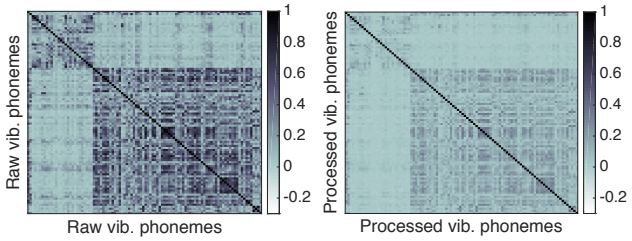

Figure 20: The heat-map shows the correlation of the frequency domain features of the phoneme sounds, recorded with custom vibration motor: (a) before processing and (b) after processing.

6.3 Performance Results with MSR

Accuracy v/s Loudness

Figure 21 shows the accuracy with manual speech recognition (MSR) in comparison to automatic (ASR). Unsurprisingly, the accuracy is around 20% more than ASR at higher loudness regimes (60 dbSPL or more) – the individuals guessed the words individually in these experiments. Using consensus from group discussion, the accuracy increases to 88% at 60 dbSPL. When the loudness is stronger, VibraPhone is comparable to microphones, both for custom hardware and smartphones.

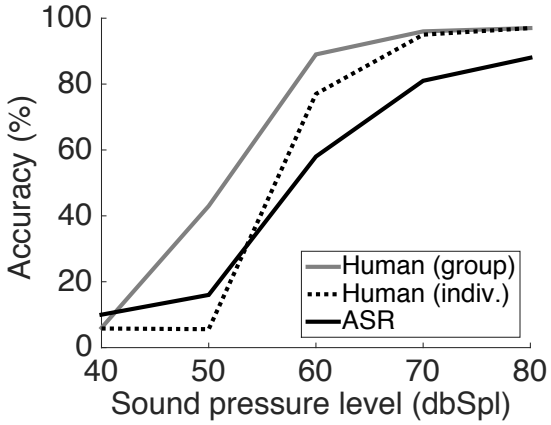

Figure 21: This plot compares the accuracy of human decoding with ASR. It shows the performance of the human decoders while working individually and as a collaborative team.

Hot-phrase Detection

Figure 22 shows manual performance with "hot phrases", i.e., where the volunteer was asked to pick a phrase from the list that best matched the spoken phrase (the volunteer could also select none of the phrases). We provided a list of 10 written phrases before playing the positive and negative samples in arbitrary sequence. Example phrases were "turn left", "happy birthday", "start the computer", etc., and the negative samples were chosen with comparable number of words and characters.

Figure 22(a) reports results from the custom hardware – volunteers almost perfectly identified the phrases and rejected the negative samples.
However, when using the smartphone vibra-motor, VibraPhone failed to identify some positive samples – Figure 22(b) shows the outcome in relatively higher false negative values. Of course, the degradation is relative – the absolute detection performance is still quite high, with accuracy and precision at 0.83 and 0.90, respectively, for the processed vibra-signal.

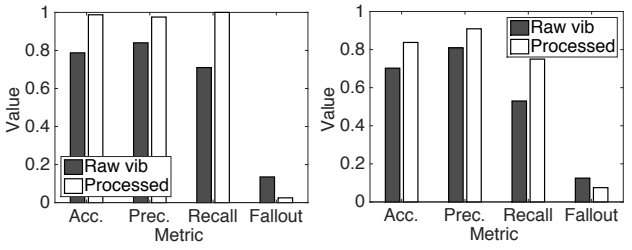

Figure 22: The accuracy, precision, recall, and fallout values for manual hot-phrase detection. The recording device is (a) the custom hardware and (b) the smartphone.

Perceived Clarity

Human volunteers also assigned a "clarity score" on a range of $[0, 10]$ to every word/phrase he/she listened to (a score of 10 indicated a perfectly intelligible word). Figure 23 plots the average clarity score of the correctly decoded samples and compares it between the vibration motor and the microphone. The subjective perception of clarity does not change for the microphone for sound pressure levels 50 dbSPL and above. While VibraPhone's clarity is lower than microphone in general, the gap reduces at higher loudness levels. At 80 dbSPL, the perceived clarity scores for microphones and VibraPhone are 9.1 and 7.6, respectively.

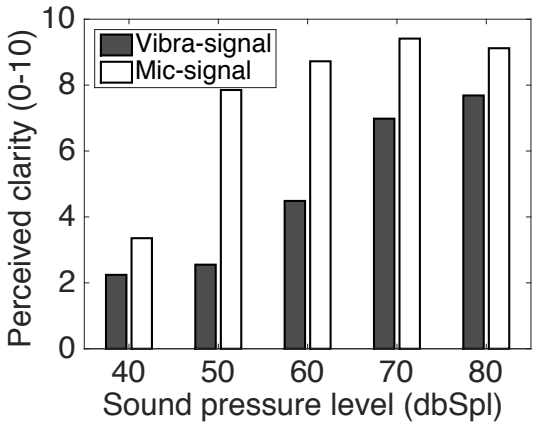

Figure 23: The perceived clarity of the correctly decoded speech recorded with microphone and vibration motor.

Kinds of Words

Figure 24 shows the top-10 and bottom-10 intelligible words from the ASR experiments. The font size is proportional to the decoding accuracy, indicating that "international" was decoded correctly most frequently, while prepositions like "a", "and", "or" were consistently missed. Unsurprisingly, longer words are decoded with higher accuracy because of better interpolation between the partially decoded phonemes. Figure 25 quantifies this with ASR and MSR, respectively – words with 5+ characters are mostly multi-syllable, yielding improved recognition.

Figure 24: Top 10 words that are (a)correctly and (b)incorrectly decoded by ASR.

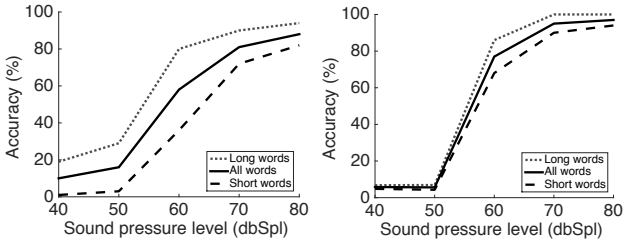

Figure 25: (a) ASR and (b) MSR accuracy for long (> 6 chars) and short (≤ 6 chars) words, as a function of loudness.

Electromagnetic Coupling:

Table 1 summarizes the manual speech recognition performance for the electromagnetic coupling test mentioned in section 2.2. In this microbenchmark we remove the equipment (microphone, speaker etc.) from the test environment that can potentially create electromagnetic coupling with the vibration motor. The signal recorded in this microbenchmark does not show any quantitative difference from that of our standard test environment. However, we run a manual speech recognition test on these recordings to identify possible perceptual differences in manual speech recognition. Here the volunteers transcribe the voice of a male non-native speaker recorded with a vibration motor during the microbenchmark test. In this test the volunteers individually listen to the recordings at sound levels according to their personal preferences. The percentage of the incorrect words in the transcription and the perceived quality score given by each user are shown in the table. The perceived sound quality is consistent with our previous results at 60db-SPL, the natural loudness of the speaker's voice at $3ft$ from the recording device.

Table 1: Coupling sensitivity data

User	A	B	C	D	E	F	G	H
Error(%)	8	0	0	8	0	0	17	25
Score	8	8	6	6	4	3	3	3

7. POINTS OF DISCUSSION

We discuss a few limitations of this paper, and a few other kinds of applications using VibraPhone.

What is the Best Possible? We have not been able to comment on the best possible performance possible with VibraPhone. Such an analysis will certainly need a deeper signal processing treatment, as well as detailed domain knowledge from speech recognition. This work is more of a lower bound on feasibility, drawing on a diverse set of established techniques from literature, and modifying them to suit the needs of this specific problem. We have initiated collaboration with signal processing researchers to push the envelope of this side channel leak.

Energy: We have sidestepped energy considerations in this paper. However, we intuitively believe that VibraPhone is not likely to be energy hungry (even though the vibra-motor consumes considerable energy while pulsating). This is because VibraPhone picks up the ambient sounds while it is in the inactive/passive mode, i.e., when it is not serving as an actuator. We plan to characterize the sensitivity and energy profile in future.

Applications: We observed that when vibra-motors are pasted to walls and floors, and music is being played in the adjacent rooms, VibraPhone is able to detect these sounds better than the microphone. We also observed that by placing the vibra-motor on the throat, various speech components can be detected, and in some cases, compliments the response of the microphone. Finally, we find that noise properties of vibra-motors and microphones are uncorrelated, enabling the possibility of diversity combining (i.e., they could together behave like a MIMO system, improving the capacity of acoustic channels). All these observations are preliminary, and hence, not reported in this paper – we plan to investigate them further as a continuation of VibraPhone.

8. RELATED WORK

Past work on acoustic side channels and speech recovery are most relevant to this paper. Given both are reasonably mature areas, we sample a subset of them.

Passive Speech Recording

Gyrophone and AccelWord [30, 47] are perhaps the closest to our work. In Gyrophone [30], authors identify the MEMS sensors' capability to capture sound. The paper presents a range of signal processing and machine learning techniques to recover traces of ambient sounds from the gyroscope data [37]. AccelWord [47] takes a step forward and uses speech information from the accelerometer [23] to implemented a low energy voice control application for a limited vocabulary of commands. However, these techniques recover only a low bandwidth of the spectrum (< 200Hz), which does not even cover the full range of fundamental frequencies in female speech ($165 - 255$ Hz). Therefore, these techniques mainly focus on extracting the reliable features of sound for consistent pattern classification. In contrast, VibraPhone concentrates on recovering a telephone-quality speech (bandwidth 4KHz [22, 31]) from the vibration motor signal, making the output amenable to manual or automatic decoding. Both Gyrophone and AccelWord are unable to produce (actually not designed for) machine understandable speech.

A family of techniques [32, 39, 45, 40] targets a light/LASER beam on an object exposed to the speech signal and records its vibration by measuring the fluctuation of the reflected beam. Visual microphone [9] is also a similar technique that uses high speed video of the target object to recover the vibration proportional to the speech signal. Camera based techniques are devoid of the noisy data that pollute motion sensors/actuators, while they must tackle other difficult challenges in computer vision. A number of solutions have monitored the change in received signal strength (RSS) and phase of the wireless radio signal reflected off the loudspeaker to capture the traces of sound. The projects [46] and [29] demonstrate successful sound recovery using reflected radio signal even when the receiver is not in the line-of-sight of the vibrating object.

Speech Recovery

We borrowed building blocks from the vast literature of speech processing. A body of research [8, 26, 25] explores artificial bandwidth expansion problems primarily to aid high quality voice transfer over band-limited telephonic channel. Some solutions attempt to identify the phonemes from the low bandwidth signal and then replace them with high bandwidth phonemes from a library. These solutions do not solve VibraPhone's problems as majority of them consider 4KHz signal as the input providing enough diversity for correct phoneme identification. VibraPhone attempts to extend the effective bandwidth from 2KHz to 4Khz – a challenge because the features up to 2KHz provide limited exposure to phonemes.

Data imputation techniques [38, 17] attempt to predict erasures in audio signals. When these signals exhibit a consistent statistical model, the erasures can be predicted well, enabling successful imputation. However, vibra-signals often lack such properties, and moreover, the location of erasures cannot be confidently demarcated.

9. CONCLUSION

This paper demonstrates that the vibration motor, present in almost all mobile devices today, can be used as a listening sensor, similar to a microphone. While this is not fundamentally surprising (since vibrating objects should respond to ambient air vibrations), the ease and extent to which the actuator has "picked up" sounds has been somewhat unexpected for us. Importantly, the decoded sounds are not merely vibration patterns that correlates to some spoken words. Rather, they actually contain the phonemes and structure of human voice, thereby requiring no machine learning or pattern recognition to extract them. We show that with basic signal processing techniques, combined with the structure of human speech, the vibra-motor's output can be quite intelligible to most human listeners. Even automatic speech recognizers were able to decode the majority of the detected words and phrases, especially at higher loudness. The application space of such systems remains open, and could range from malware eavesdropping into human phone conversation, to voice controlled wearables, to better microphones that use the vibra-motor as a second MIMO-antenna. Our ongoing work is in pursuit of a few such applications.

Acknowledgement

We sincerely thank our shepherd Dr. Xia Zhou and the anonymous reviewers for their valuable feedback. We are grateful to Qualcomm, Huawei, HP, and NSF (grant CNS-1430033 and 1423455) for partially funding this research.

10. REFERENCES

[1] Cmu sphinx. http://cmusphinx.sourceforge.net. Last accessed 6 December 2015.

[2] Sound pressure level chart. http://www.sengpielaudio.com/TableOfSoundPressureLevels.htm. Last accessed 6 December 2015.

[3] Top 10000 words from google's trillion word corpus. https://github.com/first20hours/google-10000-english. Last accessed 6 December 2015.

[4] Unit of sound pressure level. http://trace.wisc.edu/docs/2004-About-dB/. Last accessed 9 December 2015.

[5] Vibraphone project webpage. http://synrg.csl.illinois.edu/vibraphone/. Last accessed 9 December 2015.

[6] Association, I. P. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[7] Brown, A. International phonetic alphabet. *The Encyclopedia of Applied Linguistics* (2013).

[8] Chennoukh, S., Gerrits, A., Miet, G., and Sluijter, R. Speech enhancement via frequency bandwidth extension using line spectral frequencies. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on* (2001), vol. 1, IEEE, pp. 665–668.

[9] Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G. J., Durand, F., and Freeman, W. T. The visual microphone: Passive recovery of sound from video. *ACM Trans. Graph 33*, 4 (2014), 79.

[10] Eargle, J. *The Microphone Book: From mono to stereo to surround-a guide to microphone design and application*. CRC Press, 2012.

[11] Fant, G. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, vol. 2. Walter de Gruyter, 1971.

[12] Farina, A. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108* (2000), Audio Engineering Society.

[13] Fausti, P., and Farina, A. Acoustic measurements in opera houses: comparison between different techniques and equipment. *Journal of Sound and Vibration 232*, 1 (2000), 213–229.

[14] Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., and Perrett, D. I. Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour 69*, 3 (2005), 561–568.

[15] Gales, M. J. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language 12*, 2 (1998), 75–98.

[16] Galili, I., Kaplan, D., and Lehavi, Y. Teaching faradayâĂŹs law of electromagnetic induction in an introductory physics course. *American journal of physics 74*, 4 (2006), 337–343.

[17] Gemmeke, J. F., Virtanen, T., and Hurmalainen, A. Exemplar-based sparse representations for noise robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on 19*, 7 (2011), 2067–2080.

[18] Hayes, B. *Introductory phonology*, vol. 32. John Wiley & Sons, 2011.

[19] Hillenbrand, J., and Sessler, G. M. High-sensitivity piezoelectric microphones based on stacked cellular polymer films (l). *The Journal of the Acoustical Society of America 116*, 6 (2004), 3267–3270.

[20] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE 29*, 6 (2012), 82–97.

[21] Huggins-daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., and Rudnicky, A. I. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *in Proceedings of ICASSP* (2006).

[22] Jax, P., and Vary, P. On artificial bandwidth extension of telephone speech. *Signal Processing 83*, 8 (2003), 1707–1719.

[23] Johnson, C. D. Accelerometer principles. *Process Control Instrumentation Technology* (2009).

[24] Keele Jr, D. Low-frequency loudspeaker assessment by nearfield sound-pressure measurement. *Journal of the audio engineering society 22*, 3 (1974), 154–162.

[25] Kontio, J., Laaksonen, L., and Alku, P. Neural network-based artificial bandwidth expansion of speech. *Audio, Speech, and Language Processing, IEEE Transactions on 15*, 3 (2007), 873–881.

[26] Laaksonen, L., Kontio, J., and Alku, P. Artificial bandwidth expansion method to improve intelligibility and quality of amr-coded narrowband speech. In *ICASSP (1)* (2005), pp. 809–812.

[27] Ladefoged, P. The revised international phonetic alphabet. *Language* (1990), 550–552.

[28] Lapteva, O. *Speaker Perception and Recognition: An Integrative Framework for Computational Speech Processing*. kassel university press GmbH, 2011.

[29] McGrath, W. Technique and device for through-the-wall audio surveillance, Mar. 30 2005. US Patent App. 11/095,122.

[30] Michalevsky, Y., Boneh, D., and Nakibly, G. Gyrophone: Recognizing speech from gyroscope signals. In *Proc. 23rd USENIX Security Symposium (SEC?14), USENIX Association* (2014).

[31] Moreno, P. J., and Stern, R. M. Sources of degradation of speech recognition in the telephone network. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on* (1994), vol. 1, IEEE, pp. I–109.

[32] Muscatell, R. P. Laser microphone, Oct. 23 1984. US Patent 4,479,265.

[33] Nation, P., and Waring, R. Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy 14* (1997), 6–19.

[34] Ogata, K. *System dynamics*, vol. 3. Prentice Hall New Jersey, 1998.

[35] Qi, Y., and Hillman, R. E. Temporal and spectral estimations of harmonics-to-noise ratio in human voice

signals. *The Journal of the Acoustical Society of America 102*, 1 (1997), 537–543.

[36] Roy, N., Gowda, M., and Choudhury, R. R. Ripple: Communicating through physical vibration.

[37] Scarborough, J. B. *The Gyroscope: Theory and Application*. Interscience Pub., 1958.

[38] Smaragdis, P., Raj, B., and Shashanka, M. Missing data imputation for spectral audio signals. In *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on* (2009), IEEE, pp. 1–6.

[39] Smeets, G. Laser interference microphone for ultrasonics and nonlinear acoustics. *The Journal of the Acoustical Society of America 61*, 3 (1977), 872–875.

[40] Speciale, J. R. Pulsed laser microphone, Oct. 9 2001. US Patent 6,301,034.

[41] Tanner, P., Loebach, J., Cook, J., and Hallen, H. A pulsed jumping ring apparatus for demonstration of lenzâĂŹs law. *American Journal of Physics 69*, 8 (2001), 911–916.

[42] Taylor, B. *Guide for the Use of the International System of Units (SI): The Metric System*. DIANE Publishing, 1995.

[43] Titze, I. R. *Principles of voice production*. National Center for Voice and Speech, 2000.

[44] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. Phoneme recognition using time-delay neural networks. *Acoustics, Speech and Signal Processing, IEEE Transactions on 37*, 3 (1989), 328–339.

[45] Wang, C.-C., Trivedi, S., Jin, F., Swaminathan, V., Rodriguez, P., and Prasad, N. S. High sensitivity pulsed laser vibrometer and its application as a laser microphone. *Applied Physics Letters 94*, 5 (2009), 051112.

[46] Wei, T., Wang, S., Zhou, A., and Zhang, X. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking* (2015), ACM, pp. 130–141.

[47] Zhang, L., Pathak, P. H., Wu, M., Zhao, Y., and Mohapatra, P. Accelword: Energy efficient hotword detection through accelerometer. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services* (2015), ACM, pp. 301–315.