The visual microphone: Passive recovery of sound from video

Citation: Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Fredo Durand, and William T. Freeman. 2014. The visual microphone: passive recovery of sound from video. ACM Trans. Graph. 33, 4, Article 79 (July 2014), 10 pages.

As Published: http://dx.doi.org/10.1145/2601097.2601119

Publisher: Association for Computing Machinery (ACM)

Persistent URL: http://hdl.handle.net/1721.1/100023

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

The Visual Microphone: Passive Recovery of Sound from Video

Abe Davis[1] Michael Rubinstein[2,1] Neal Wadhwa[1] Gautham J. Mysore[3] Frédo Durand[1] William T. Freeman[1]

[1]MIT CSAIL [2]Microsoft Research [3]Adobe Research

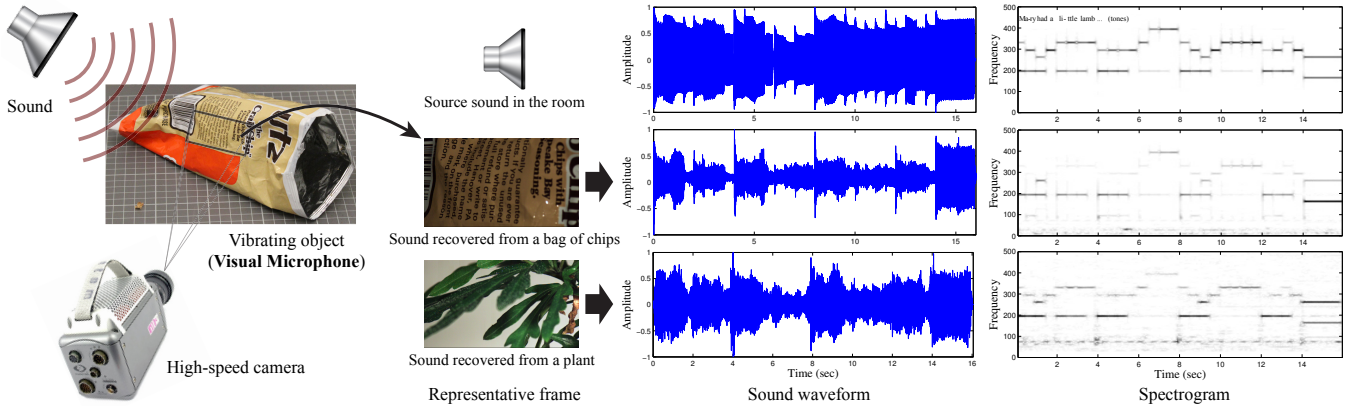


Figure 1: *Recovering sound from video. Left: when sound hits an object (in this case, an empty bag of chips) it causes extremely small surface vibrations in that object. We are able to extract these small vibrations from high speed video and reconstruct the sound that produced them - using the object as a visual microphone from a distance. Right: an instrumental recording of "Mary Had a Little Lamb" (top row) is played through a loudspeaker, then recovered from video of different objects: a bag of chips (middle row), and the leaves of a potted plant (bottom row). For the source and each recovered sound we show the waveform and spectrogram (the magnitude of the signal across different frequencies over time, shown in linear scale with darker colors representing higher energy). The input and recovered sounds for all of the experiments in the paper can be found on the project web page.*

Abstract

When sound hits an object, it causes small vibrations of the object's surface. We show how, using only high-speed video of the object, we can extract those minute vibrations and partially recover the sound that produced them, allowing us to turn everyday objects—a glass of water, a potted plant, a box of tissues, or a bag of chips—into visual microphones. We recover sounds from high-speed footage of a variety of objects with different properties, and use both real and simulated data to examine some of the factors that affect our ability to visually recover sound. We evaluate the quality of recovered sounds using intelligibility and SNR metrics and provide input and recovered audio samples for direct comparison. We also explore how to leverage the rolling shutter in regular consumer cameras to recover audio from standard frame-rate videos, and use the spatial resolution of our method to visualize how sound-related vibrations vary over an object's surface, which we can use to recover the vibration modes of an object.

CR Categories: I.4.7 [Image Processing and Computer Vision]: Scene Analysis—Time-varying Imagery;

Keywords: remote sound acquisition, sound from video, visual acoustics

Links: DL PDF WEB

1 Introduction

Sound waves are fluctuations in pressure that travel through a medium. When sound hits an object, it causes the surface of that object to move. Depending on various conditions, the surface may move with the surrounding medium or deform according to its vibration modes. In both cases, the pattern of motion contains useful information that can be used to recover sound or learn about the object's structure.

Vibrations in objects due to sound have been used in recent years for remote sound acquisition, which has important applications in surveillance and security, such as eavesdropping on a conversation from afar. Existing approaches to acquire sound from surface vibrations at a distance are *active* in nature, requiring a laser beam or pattern to be projected onto the vibrating surface.

A key observation in our work is that the vibrations that sound causes in an object often create enough *visual* signal to partially recover the sounds that produced them, using only a high-speed video of the object. Remarkably, it is possible to recover comprehensible speech and music in a room from just a video of a bag of chips (Figure 1, Figure 2).

Following this observation, we propose a *passive* method to recover audio signals using video. Our method visually detects small vibrations in an object responding to sound, and converts those vibrations back into an audio signal, turning visible everyday objects into potential microphones. To recover sound from an object, we film the object using a high-speed video camera. We then extract local motion signals across the dimensions of a complex steerable pyramid built on the recorded video. These local signals are aligned and averaged into a single, 1D motion signal that captures global movement of the object over time, which we further filter and denoise to produce the recovered sound.

Our method typically does not recover sound as well as active techniques for sound and vibration measurement, but it does provide a few advantages. In particular, it does not require active lighting for

textured objects and well-illuminated scenes (Figure 2), and does not rely on additional sensors or detection modules other than a high-speed video camera. It also does not require that the vibrating surface be retroreflective or specular (as is often required by laser microphones), and does not impose significant constraints on the surface orientation with respect to the camera. Moreover, since our method produces a spatial measurement of the sound (an estimated audio signal at every pixel in the video), we can use it to analyze sound-induced deformations of an object.

While sound can travel through most matter, not all objects and materials are equally good for visual sound recovery. The propagation of sound waves in a material depends on various factors, such as the density and compressibility of the material, as well as the object's shape. We performed controlled experiments where we measured the responses of different objects and materials to known and unknown sounds, and evaluated our ability to recover these sounds from high-speed video using our technique.

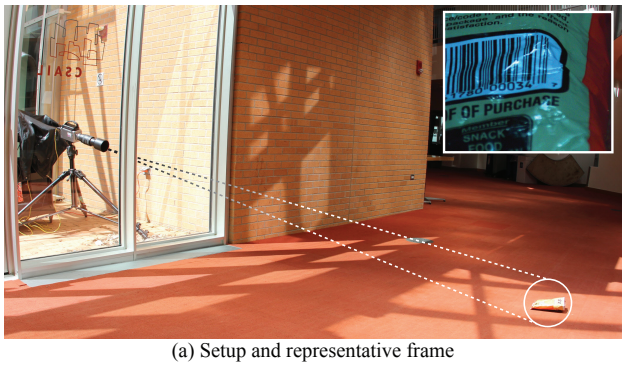
We first describe our technique in detail and show results on a variety of different objects and sounds. We then characterize the behavior and limits of our technique by looking at data from calibrated experiments and simulations. Finally, we exploit the rolling shutter of CMOS sensors to show how sound may be recovered using regular consumer cameras with standard frame-rates.

2 Related Work

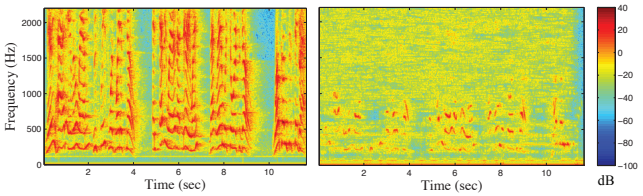
Traditional microphones work by converting the motion of an internal diaphragm into an electrical signal. The diaphragm is designed to move readily with sound pressure so that its motion can be recorded and interpreted as audio. Laser microphones work on a similar principle but instead measure the motion of a distant object, essentially using the object as an external diaphragm. This is done by recording the reflection of a laser pointed at the object's surface. The most basic type of laser microphone records the phase of the reflected laser, which gives the object's distance modulo the laser's wavelength. A laser Doppler vibrometer (LDV) resolves the ambiguity of phase wrapping by measuring the Doppler shift of the reflected laser to determine the velocity of the reflecting surface [Rothberg et al. 1989]. Both types of laser microphone can recover high quality audio from a great distance, but depend on precise positioning of a laser and receiver relative to a surface with appropriate reflectance.

Zalevsky et al. [2009] address some of these limitations by using an out-of-focus high-speed camera to record changes in the speckle pattern of reflected laser light. Their work allows for greater flexibility in the positioning of a receiver, but still depends on recording reflected laser light. In contrast, our technique does not depend on active illumination.

As our approach relies on the ability to extract extremely subtle motions from video, it is also related to recent work on magnifying and visualizing such motions [Wu et al. 2012; Wadhwa et al. 2013; Wadhwa et al. 2014; Rubinstein 2014]. These works focus on visualizing small motions, while in this paper we focus on measuring such motions and using them to recover sound. The local motion signals used in our work are derived from phase variations in the complex steerable pyramid proposed by Simoncelli et al. [1992], since these variations were shown to be well-suited for recovering small motions in video [Wadhwa et al. 2013]. However, it is also possible to compute the local motion signals using other techniques. For example, classical optical flow and point correlation methods were successfully used in previous work on visual vibration sensing [Morlier et al. 2007; D'Emilia et al. 2013]. In our case, as our output is a 1D motion signal for a single vibrating object, we are able to average over all pixels in an input video and handle extremely subtle motions, on the order of one thousandth of a pixel.



(a) Setup and representative frame



(b) Input sound (c) Recovered sound

Figure 2: *Speech recovered from a 4 kHz video of a bag of chips filmed through soundproof glass. The chip bag (on the floor on the bottom right in (a)) is lit by natural sunlight only. The camera (on the left in (a)) is positioned outside the room behind thick soundproof glass. A single frame from the recorded video (400 × 480 pixels) is shown in the inset. The speech "Mary had a little lamb ... Welcome to SIGGRAPH!" was spoken by a person near the bag of chips. (b) and (c) show the spectrogram of the source sound recorded by a standard microphone next to the chip bag, and the spectrogram of our recovered sound, respectively. The recovered sound is noisy but comprehensible (the audio clips are available on the project web page).*

3 Recovering Sound from Video

Figure 3 gives a high-level overview of how the visual microphone works. An input sound (the signal we want to recover) consists of fluctuations in air pressure at the surface of some object. These fluctuations cause the object to move, resulting in a pattern of displacement over time that we film with a camera. We then process the recorded video with our algorithm to recover an output sound.

The input to our method is a video, $V(x, y, t)$, of an object. In this section we consider high speed video (1kHz-20kHz). Lower frame rates are discussed in Section 6. We assume that the relative motion of our object and camera is dominated by vibrations due to a sound signal, $s(t)$. Our goal is to recover $s(t)$ from V.

We proceed in three steps. First, we decompose the input video V into spatial subbands corresponding to different orientations θ and scales r. We then compute local motion signals at every pixel, orientation, and scale. We combine these motion signals through a sequence of averaging and alignment operations to produce a single global motion signal for the object. Finally, we apply audio denoising and filtering techniques to the object's motion signal to obtain our recovered sound.

3.1 Computing Local Motion Signals

We use phase variations in a complex steerable pyramid representation of the video V to compute local motion.The complex steerable pyramid [Simoncelli et al. 1992; Portilla and Simoncelli 2000] is a filter bank that breaks each frame of the video $V(x, y, t)$ into complex-valued sub-bands corresponding to different scales and orientations. The basis functions of this transformation are scaled

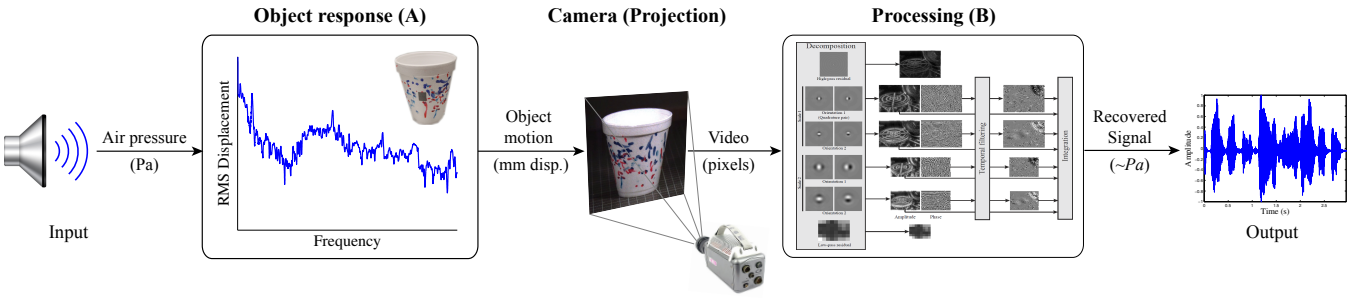


Figure 3: *We model the visual microphone as a system that operates on sound. Component* **A** *(Section 5.1) models an object's response to sound, and is purely physical—taking as input changes in air pressure, measured in Pascals, and producing physical displacement of the object over time, measured in millimeters. The response of the object to the sound depends on various factors such as the sound level at the object, and the object's material and shape. A camera then records the object, transforming the physical displacements into pixel motions in a video. Component* **B** *(Section 3, Section 5.2) is our spatiotemporal processing pipeline, which transforms the motions in the video back into sound. The resulting 1D signal is unit-less, but is correlated with the input Pascals and can therefore be played and analyzed as sound.*

and oriented Gabor-like wavelets with both cosine- and sine-phase components. Each pair of cosine- and sine-like filters can be used to separate the amplitude of local wavelets from their phase. Specifically, each scale r and orientation θ is a complex image that can be expressed in terms of amplitude A and phase φ as

$$A(r, \theta, x, y, t)e^{i\varphi(r,\theta,x,y,t)}. \tag{1}$$

We take the local phases φ computed in this equation and subtract them from the local phase of a reference frame t_0 (typically the first frame of the video) to compute the phase variations

$$\varphi_v(r, \theta, x, y, t) = \varphi(r, \theta, x, y, t) - \varphi(r, \theta, x, y, t_0). \tag{2}$$

For small motions, these phase variations are approximately proportional to displacements of image structures along the corresponding orientation and scale [Gautama and Van Hulle 2002].

3.2 Computing the Global Motion Signal

For each orientation θ and scale r in the complex steerable pyramid decomposition of the video, we compute a spatially weighted average of the local motion signals to produce a single motion signal $\Phi(r, \theta, t)$. We perform a weighted average because local phase is ambiguous in regions that do not have much texture, and as a result motion signals in these regions are noisy. The complex steerable pyramid amplitude A gives a measure of texture strength, and so we weigh each local signal by its (squared) amplitude:

$$\Phi_i(r, \theta, t) = \sum_{x,y} A(r, \theta, x, y)^2 \varphi_v(r, \theta, x, y, t). \tag{3}$$

Before averaging the $\Phi(r, \theta, t)$ over different scales and orientations, we align them temporally in order to prevent destructive interference. To understand why we do this, consider the case where we want to combine just two orientations (x and y) from a single spatial scale. Now, consider a small Gaussian vibrating in the direction $y = -x$. Here, changes in the phases of our x and y orientation will be negatively correlated, always summing to a constant signal. However, if we align the two phase signals (by shifting one of them in time) we can cause the phases to add constructively. The aligned signals are given by $\Phi(r_i, \theta_i, t - t_i)$, such that

$$t_i = \underset{t_i}{\operatorname{argmax}}\, \Phi_0(r_0, \theta_0, t)^T \Phi_i(r_i, \theta_i, t - t_i), \tag{4}$$

where i indexes all scale-orientation pairs (r, θ), and $\Phi_0(r_0, \theta_0, t)$ is an arbitrary choice of reference scale and orientation. A similar correlation metric was used by [Liu et al. 2005] to cluster related motions for motion magnification.

Our global motion signal is then:

$$\hat{s}(t) = \sum_i \Phi_i(r_i, \theta_i, t - t_i), \tag{5}$$

which we scale and center to the range $[-1, 1]$.

3.3 Denoising

We further process the recovered global motion signal to improve its SNR. In many videos, we noticed high energy noise in the lower frequencies that typically did not correspond to audio. We address this by applying a high pass Butterworth filter with a cutoff of 20-100Hz (for most examples, $1/20$ of the Nyquist frequency)[1].

Our choice of algorithm for additional denoising depends on our target application – specifically, whether we are concerned with accuracy or intelligibility. For applications targeting accuracy we use our own implementation of a technique known as spectral subtraction [Boll 1979]. For intelligibility we use a perceptually motivated speech enhancement algorithm [Loizou 2005] that works by computing a Bayesian optimal estimate of the denoised signal with a cost function that takes into account human perception of speech. All of the results we present in this paper were denoised automatically with one of these two algorithms. Our results may be further improved by using more sophisticated audio denoising algorithms available in professional audio processing software (some of which require manual interaction).

Different frequencies of our recovered signal might be modulated differently by the recorded object. In section 4.3, we show how to use a known test signal to characterize how an object attenuates different frequencies, then use this information to equalize unknown signals recovered from the same object (or a similar one) in new videos.

4 Experiments

We performed a variety of experiments to test our technique. All the videos in this section were recorded indoors with a Phantom V10 high speed camera. The setup for these experiments consisted of an object, a loudspeaker, and the camera, arranged as shown in Figure 4. The loudspeaker was always placed on its own stand separate from the surface holding the object in order to avoid contact

[1]For very noisy cases we instead apply this highpass filter to the $\Phi(r, \theta, t)$ signals before alignment to prevent the noise from affecting the alignment.

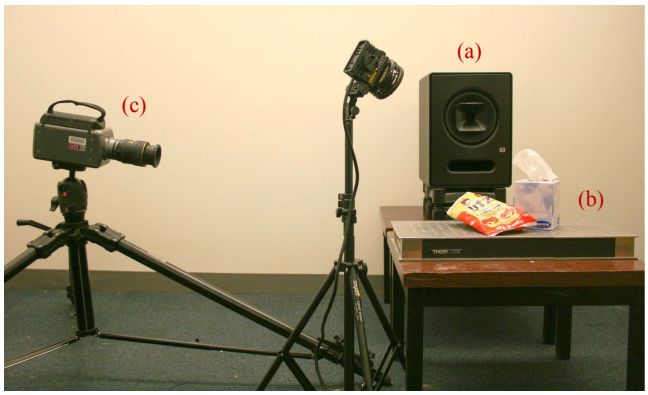


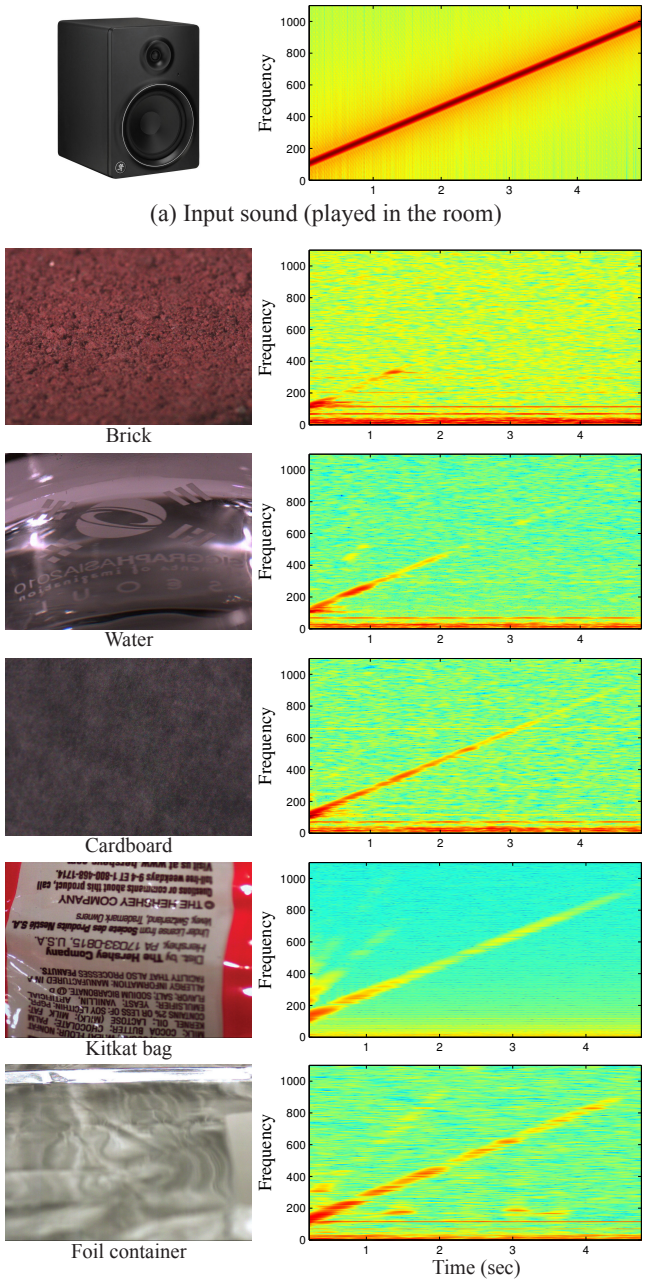
Figure 4: *An example of our controlled experimental setup. Sound from an audio source, such as a loudspeaker (a) excites an ordinary object (b). A high-speed camera (c) records the object. We then recover sound from the recorded video. In order to minimize undesired vibrations, the objects were placed on a heavy optical plate, and for experiments involving a loudspeaker we placed the loudspeaker on a separate surface from the one containing the objects, on top of an acoustic isolator.*

vibrations. The objects were lit with photography lamps and filmed at distances ranging from 0.5 meter to 2 meters. In other experiments we recover sound from greater distances without the aid of photography lamps (e.g. Figure 2). Video frame rates are in the range of 2kHz-20kHz, with resolutions ranging from 192x192 pixels to 700x700 pixels. Sounds were played at loud volumes ranging from 80 dB (an actor's stage voice) to 110 dB (comparable to a jet engine at 100 meter). Lower volumes are explored in Section 5, Figure 2, and additional experiments on our web page. Videos were processed using complex steerable pyramids with 4 scales and 2 orientations, which we computed using the publicly available code of Portilla and Simoncelli [2000]. Processing each video typically took 2 to 3 hours using MATLAB on a machine with two 3.46GHz processors and 32GB of RAM.

Our first set of experiments tested the range of frequencies that could be recovered from different objects. We did this by playing a linear ramp of frequencies through the loudspeaker, then seeing which frequencies could be recovered by our technique. The second set of experiments focused on recovering human speech from video. For these experiments we used several standard speech examples from the TIMIT dataset [Fisher et al. 1986] played through a loudspeaker, as well as live speech from a human subject (here the loudspeaker in Figure 4 was replaced with a talking human). Audio for these experiments and others can be found on the project website. Our results are best experienced by listening to the accompanying audio files through headphones.

4.1 Sound Recovery from Different Objects/Materials

In this first set of experiments we play a ramp signal, consisting of a sine wave that increases linearly in frequency over time, at a variety of objects. Figure 5(a) shows the spectrogram of our input sound, which increases from 100Hz to 1000Hz over 5 seconds. Figure 5(b) shows the spectrograms of signals recovered from 2.2kHz videos of a variety of objects with different material properties. The brick at the top of Figure 5(b) is used as a control experiment where we expect to recover little signal because the object is rigid and heavy. The low-frequency signal recovered from the brick (see the spectrogram visualized for *Brick* in Figure 5(b)) may come from motion of the brick or the camera, but the fact that this signal is very weak suggests that camera motion and other unintended factors in the experimental setup have at most a minor impact on our results. In



(a) Input sound (played in the room)

(b) Reconstructed sound

Figure 5: *Sound reconstructed from different objects and materials. A linear ramp ranging from* $100 - 1000Hz$ *was played through a loudspeaker (a), and reconstructed from different objects and materials (b). In* Water*, the camera was pointed at one side of a clear mug containing water, where the water surface was just above a logo printed on the side of the mug. Motion of the water's surface resulted in changing refraction and moving specular reflections. More details can be found on our project web page.*

particular, while almost no signal is recovered from the brick, much better signal is recovered from the other objects shown.

In almost all of our results the recovered signal is weaker in higher frequencies. This is expected, as higher frequencies produce smaller displacements and are attenuated more heavily by most materials. We show this more explicitly with data from a laser Doppler vibrometer in Section 5. However, the decrease in power with

Sequence	Method	SSNR	LLR Mean	Intelligibility
Female speaker - fadg0, sa1	VM	24.5	1.47	0.72
	LDV	**28.5**	**1.81**	**0.74**
Female speaker - fadg0, sa2	VM	**28.7**	1.37	0.65
	LDV	26.5	**1.82**	**0.70**
Male speaker - mccs0, sa1	VM	20.4	1.31	0.59
	LDV	**26.1**	**1.83**	**0.73**
Male speaker - mccs0, sa2	VM	23.2	1.55	0.67
	LDV	**25.8**	**1.96**	**0.68**
Male speaker - mabw0, sa1	VM	23.3	1.68	**0.77**
	LDV	**28.2**	**1.74**	0.76
Male Speaker - mabw0, sa2	VM	25.5	1.81	0.72
	LDV	**26.0**	**1.88**	**0.74**

Table 1: *A comparison of our method (VM) with a laser Doppler vibrometer (LDV). Speech from the TIMIT dataset is recovered from a bag of chips by both methods simultaneously. Both recovered signals are denoised using [Loizou 2005]. The recovered signals are evaluated using Segmental SNR (SSNR, in dB) [Hansen and Pellom 1998], Log Likelihood Ratio mean (LLR) [Quackenbush et al. 1988] and the intelligibility metric described in [Taal et al. 2011] (given in the range* 0-1*). For each comparison, the better score is shown in bold.*

higher frequencies is not monotonic, possibly due to the excitement of vibration modes. Not surprisingly, lighter objects that are easier to move tend to support the recovery of higher frequencies better than more inert objects.

4.2 Speech Recovery

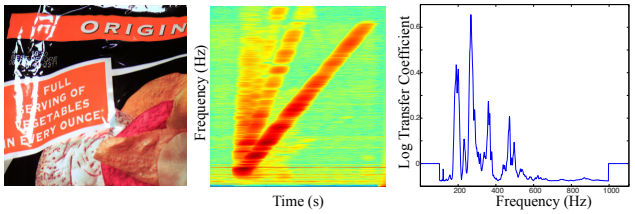
Speech recovery is an exciting application of the visual microphone. To test our ability to recover speech we use standard speech examples from the TIMIT dataset [Fisher et al. 1986], as well as live speech from a human speaker reciting the poem "Mary had a little lamb," in reference to the first words spoken by Thomas A. Edison into the Phonograph in 1877. Additional speech experiments can be found on the project website.

In most of our speech recovery experiments, we filmed a bag of chips at 2200 FPS with a spatial resolution of 700×700 pixels. Recovered signals were denoised with a perceptually motivated speech enhancement algorithm [Loizou 2005], described in section 3.3.

The best way to evaluate our reconstructed speech is to listen to the accompanying audio files, available on our project website. In addition to providing these audio files, we also evaluate our results using quantitative metrics from the audio processing community. To measure accuracy we use Segmental Signal-to-Noise Ratio (SSNR) [Hansen and Pellom 1998], which averages local SNR over time. To measure intelligibility we use the perceptually-based metric of Taal et al. [2011]. For our results in Table 1 we also include Log Likelihood Ratio (LLR) [Quackenbush et al. 1988], which is a metric that captures how closely the spectral shape of a recovered signal matches that of the original clean signal. Finally, our results can be evaluated visually by looking at the spectrograms of our input speech and recovered signals, shown in Figure 6.

Up to the Nyquist frequency of our videos, the recovered signals closely match the input for both pre-recorded and live speech. In one experiment, we captured a bag of chips at 20,000 FPS and were able to recover some of the higher frequencies of the speech (Figure 6, bottom right). The higher frame rate resulted in reduced exposure time and therefore more image noise, which is why the resulting figure is noisier than the results at 2200Hz. However, even with this added noise, we were able to qualitatively understand the speech in the reconstructed audio.

We also compare our results to audio recovered by a laser Doppler vibrometer (Table 1). Our method recovered audio that was comparable to the laser vibrometer when sampled at the same rate as the video, as measured by the intelligibility metric. However, the LDV required active lighting, and we had to affix a piece of retro-reflective tape on the object for the laser to bounce off the object and go back to the vibrometer. Without the retro-reflective tape, the quality of the vibrometer signal was significantly worse.



(a) Input (b) Spectrogram (c) Log Transfer coefficients

Speaker	fadg0		mccs0		mabw0	
Clip	sa1	sa2	sa1	sa2	sa1	sa2
SSNR w/o Eq.	33.2	29.7	29.8	30.4	19.6	**30.7**
SSNR with Eq.	**35.9**	**33.2**	**30.1**	**31.8**	**20.9**	27.8

Table 2: *We use a known ramp signal to estimate the transfer coefficients for a bag of chips. We then use these transfer coefficients to equalize new unknown signals recovered from the same bag. a) One frame from a video of the bag of chips. b) The recovered ramp signal we use to compute transfer coefficients. c) The log transfer coefficients (set to 1 outside the range of frequencies in our ramp). The table shows SSNR for six speech examples with and without the equalization. Spectral subtraction is applied again after equalization, as boosting attenuated frequencies tends to boost noise in those frequencies as well. Note that the denoising method SSNR values reported here are different from Table 1, as our equalization focuses on accuracy over intelligibility (see text for details).*

4.3 Transfer Functions and Equalization

We can use the ramp signal from Section 4.1 to characterize the (visual) frequency response of an object in order to improve the quality of signals recovered from new observations of that object. In theory, if we think of the object as a linear system, Wiener deconvolution can be used to estimate the complex-valued transfer function associated with that system, and that transfer function could then be used to deconvolve new observed signals in an optimal way (in the mean squared error sense). In practice however, this approach can be highly susceptible to noise and nonlinear artifacts. Instead, we describe a simpler method that first uses the short time Fourier transform of a training example (the linear ramp) to calculate frequency transfer coefficients at a coarse scale, then equalizes new observed signals using these transfer coefficients.

Our transfer coefficients are derived from the short time power spectra of an input/output pair of signals (like the ones shown in Figure 5). Each coefficient corresponds to a frequency in the short time power spectra of the observed training signal, and is computed as a weighted average of that frequency's magnitude over time. The weight at every time is given by the short time power spectrum of the aligned input training signal. Given that our input signal contains only one frequency at a time, this weighting scheme ignores nonlinear artifacts such the frequency doubling seen in Figure 2(b).

Once we have our transfer coefficients we can use them to equalize new signals. There are many possible ways to do this. We apply gains to frequencies in the short time power spectra of the new signal, then resynthesize the signal in the time domain. The gain we apply to each frequency is proportional to the inverse of its corresponding transfer coefficient raised to some exponent k.

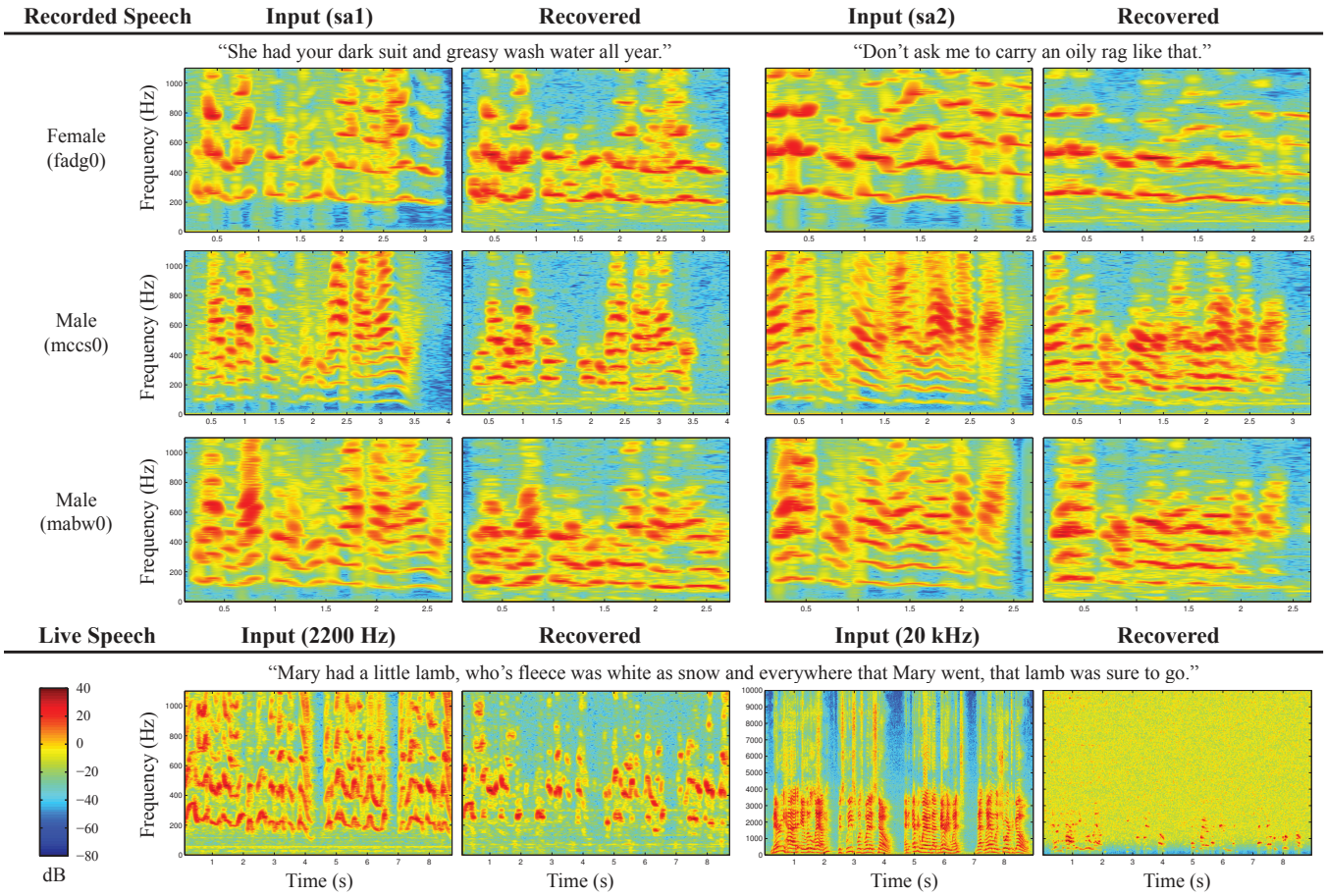


Figure 6: *Speech recovered from a bag of chips.* ***Recorded Speech (top three rows):*** *We play recordings of three speakers saying two different sentences from the TIMIT dataset [Fisher et al. 1986] through a loudspeaker near a bag of chips. We then recover audio from a* $2,200$*Hz, 700×700 video of the bag of chips (see table 2(a)) for a representative frame) and display the spectrograms of both the input audio and the recovered signal.* ***Live Speech (bottom row):*** *In a separate experiment, a male speaker recites the nursery rhyme "Mary had a little lamb...", near the same bag of chips. We display the spectrograms of audio recorded by a conventional microphone next to the spectrograms of the audio recovered from video of the bag of chips using our technique. Results were recovered from videos taken at* $2,200$*Hz, 700× 700 pixels (bottom left), and 20 kHz, 192× 192 pixels (bottom right). Input and recovered audio clips can be found on the project web page.*

Figure 2 shows the results of applying an equalizer derived from a chip bag to speech sequences recovered from the same object. In the absence of noise, k would be set to 1, but broad spectrum noise compresses the range of the estimated transfer coefficients. Using a larger k can compensate for this. We manually tuned k on one of the female speech examples, then applied the resulting equalizer to all six speech examples. Since this equalization is designed to improve the faithfulness of a recovered signal rather than the intelligibility of speech, we use spectral subtraction for denoising and SSNR to evaluate our results.

Note that calibration and equalization are optional. In particular, all of the results in this paper outside of Table 2 assume no prior knowledge of the recorded object's frequency response.

5 Analysis

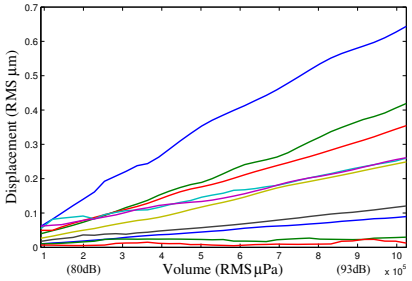
In this section, we provide an analysis that helps predict when and how well our technique works, and estimate the scale of motions that we are able to recover. At a high level, our method tries to infer some input sound $s(t)$ by observing the motion it causes in a nearby object. Figure 3 outlines a series of transformations describing this process. A sound, $s(t)$, defined by fluctuations in air pressure over time, acts on the surface of an object. The object then moves in response to this sound, transforming air pressure into surface displacement. We call this transformation the object response, $\mathbf{A}$. The resulting pattern of surface displacement is then recorded with a camera, and our algorithm, $\mathbf{B}$, transforms the recorded video into a recovered sound. Intuitively, our ability to recover $s(t)$ will depend on the transformations $\mathbf{A}$ and $\mathbf{B}$. In this section we characterize these transformations to help predict how well the visual microphone will work in new situations.

5.1 Object Response ($\mathbf{A}$)

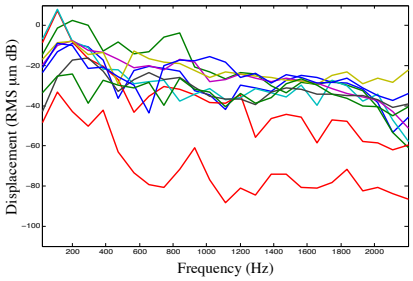
For each object we recorded motion in response to two signals in a calibrated lab setting. The first was a 300Hz pure tone that increased linearly in volume from [0.1-1] Pascals (RMS) (˜57 to 95 decibels). This signal was used to characterize the relationship between volume and object motion. To get an accurate measure of volume we calibrated our experimental setup (the loudspeaker, room, and position of the object being tested) using a decibel meter. Figure 7 (b) shows the RMS motion of different objects as a function of RMS air pressure in Pascals (at 300Hz). From this graph we see that for most of the objects we tested, the motion appears to be approximately linear in sound pressure. For each object we tested

	Measured displacement increase rate (ε m/Pa)
crabchips	62.4
greenteabox	41.1
tissue	32.9
foiltogo	20.6
kitkat	21.6
afoil	24.6
rose	10.6
foamcup	8.6
chobani	1.2
teapot	1.1

(a) Displacement coefficients at 300Hz



(b) Motion vs. sound volume



(c) Frequency responses

Figure 7: *Object motion as function of sound volume and frequency, as measured with a laser Doppler vibrometer. Top: the objects we measured, ordered according to their peak displacement at 95 dB, from left (larger motion) to right (smaller motion). (b) The RMS displacement (micrometers) vs RMS sound pressure (Pascals) for the objects being hit by a calibrated 300Hz sine wave linearly increasing in volume from 57 decibels to 95 decibels. Displacements are approximately linear in Pascals, and are all in the order of a micrometer (one thousandths of a millimeter). (c) The frequency responses of these objects (Power dB vs frequency), based on their response to a ramp of frequencies ranging from 20Hz to 2200Hz. Higher frequencies tend to have weaker responses than lower frequencies. Frequency responses are plotted on a dB scale, so the relative attenuation of higher frequencies is quite significant.*

one or more additional frequencies and saw that this relationship remained linear, suggesting that we may model the object response $\mathbf{A}$ as a linear time invariant (LTI) system.

Our second test signal was a ramp signal similar to the one used in Section 4.1, with frequencies in the range of 20Hz to 2200Hz. Modeling $\mathbf{A}$ as an LTI system, we used this ramp signal to recover the impulse response of that system. This was done by deconvolving our observed ramp signal (this time recorded by a LDV) by our known input using Wiener deconvolution. Figure 7 (c) shows frequency responses derived from our recovered impulse responses[2]. From this graph we see that most objects have a stronger response at lower frequencies than higher frequencies (as expected), but that this trend is not monotonic. This agrees with what we observed in Section 4.1.

We can now express the transformation $\mathbf{A}$ in the frequency domain as multiplication of our sound spectrum, $S(\omega)$, by the transfer function $\mathbf{A}(\omega)$, giving us the spectrum of our motion, $D_{mm}(\omega)$:

$$D_{mm}(\omega) \approx \mathbf{A}(\omega)S(\omega) \tag{6}$$

The magnitude of the coefficient $\mathbf{A}(\omega)$ for an object corresponds to the slope of its respective volume vs. displacement curve (like the ones shown in Figure 7(b)) at frequency ω.

5.2 Processing (B)

The relationship between object motion D_{mm} and pixel displacement, D_p, is a straightforward one given by the projection and sampling of a camera. Camera parameters like distance, zoom, viewing angle, etc., affect our algorithm's input (the video) by changing the number of pixels that see an object, n_p, the magnification of pixel motion (in $\text{mm}/pixel$), m, and the noise of captured images, σ_N.

[2]The frequency responses shown here have been smoothed to remove noise and intelligibly display all ten on one graph. Responses may also be affected by the responses of the room and speaker.

The relationship between object motion and pixel motion can be expressed as:

$$D_p(\omega) = D_{mm}(\omega) \times m \times \cos(\theta) \tag{7}$$

where θ is the viewing angle of our camera relative to the object's surface motion and m is the magnification of our surface in $\frac{\text{mm}}{pixel}$.

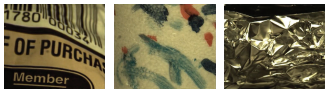
Through simulations we also studied the effect of the number of pixels imaging an object (n_p), the amplitude (in pixels) of motion ($D_p(\omega)$), and image noise (given by standard deviation σ_n), on the SNR of our recovered sounds. The results of these simulations (available on our webpage) confirmed the following relationship:

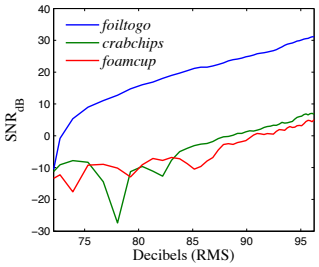
$$\frac{\sigma_S(\omega)}{\sigma_N(\omega)} \propto |D_p(\omega)| \frac{\sqrt{n_p}}{\sigma_n}, \tag{8}$$

which shows how the signal to noise ratio increases with motion amplitude and the number of pixels, and decreases with image noise.

To confirm this relationship between SNR and motion amplitude with real data and to test the limits of our technique on different objects, we conducted another calibrated experiment like the one discussed in Section 5.1, this time using the visual microphone instead of a laser vibrometer. In this experiment, the camera was placed about 2 meters away from the object being recorded and objects were imaged at 400×480 pixels with a magnification of 17.8 pixels per millimeter. With this setup, we evaluated SNR (dB) as a function of volume (standard decibels). For sufficiently large amplitudes of pixel displacement, our recovered signal becomes approximately linear in volume (Fig. 8(a)), confirming the relationship given in Equation 8.

To give a sense of the size of motions in our videos, we also estimated the motion, in pixels, for each of the of the corresponding videos using phase-based optical flow [Gautama and Van Hulle 2002]. We found these motions to be on the order of one hundredth to one thousandth of a pixel (Fig. 8(b)).





(a) SNR vs. Volume

Object	Freq. (Hz)	RMS (px)
crabchips	100	0.029
foiltogo	100	0.010
foamcup	100	0.007
crabchips	300	0.006
foiltogo	300	0.012
foamcup	300	0.005
crabchips	500	0.005
foiltogo	500	0.007
foamcup	500	0.002

(b) Motion in Pixels

Figure 8: *The signal-to-noise ratio of sound recovered from video as a function of volume (a), and the absolute motion in pixels (b), for several objects when a sine wave of varying frequency and volume is played at them.*

6 Recovering Sound with Normal Video Cameras using Rolling Shutter

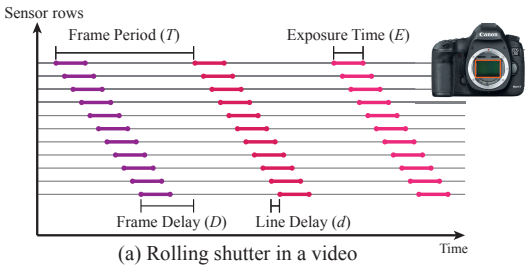
One limitation of the technique presented so far is the need for high speed video. We explore the possibility of recovering audio from video filmed at regular frame rates by taking advantage of the *rolling shutter* common in the CMOS sensors of most cell phones and DSLR cameras [Nakamura 2005]. With rolling shutter, sensor pixels are exposed and read out row-by-row sequentially at different times from top to bottom. Compared to uniform global shutters, this design is cheaper to implement and has lower power consumption, but often produces undesirable skewing artifacts in recorded images, especially for photographs of moving objects. Previously, researchers have tried to mitigate the effect of rolling shutter on computer vision problems such as structure-from-motion [Meingast et al. 2005] and video stabilization [Grundmann et al. 2012]. Ait-Aider et al. [2007] used rolling shutter to estimate the pose and velocity of rigid objects from a single image. We take advantage of rolling shutter to effectively increase the sampling rate of a camera and recover sound frequencies above the camera's frame rate.

Because each row in a sensor with rolling sensor is captured at different times, we can recover an audio signal for each row, rather than each frame, increasing the sampling rate from the frame rate of the camera to the rate at which rows are recorded (Fig. 9). We can fully determine the mapping of the sensor rows to the audio signal by knowing the exposure time of the camera, E, the line delay, d, which is the time between row captures, the frame period T, the time between frame captures, and the frame delay, D (Fig. 9). The rolling shutter parameters can be taken from the camera and sensor specs, or computed (for any camera) through a simple calibration process [Meingast et al. 2005], which we also describe on our project web page. We further assume a forward model in which an object, whose image is given by $B(x, y)$, moves with coherent fronto-parallel horizontal motion described by $s(t)$, and that the motion reflects the audio we want to recover, as before. If we assume that the exposure time $E \approx 0$, then the nth frame I_n taken by the camera can be characterized by the equation

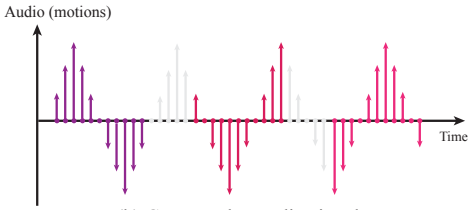
$$I_n(x, y) = B(x - \alpha s(nT + yd), y). \quad (9)$$

We use this equation to produce a simulation of rolling shutter.

If we assume that the yth row of B has sufficient horizontal texture, we can recover $s(nT + yd)$ using phase-based motion analysis. If the frame delay, the time between the capture of the last row of one frame and the first row of the next frame, is not zero, then there are



(a) Rolling shutter in a video



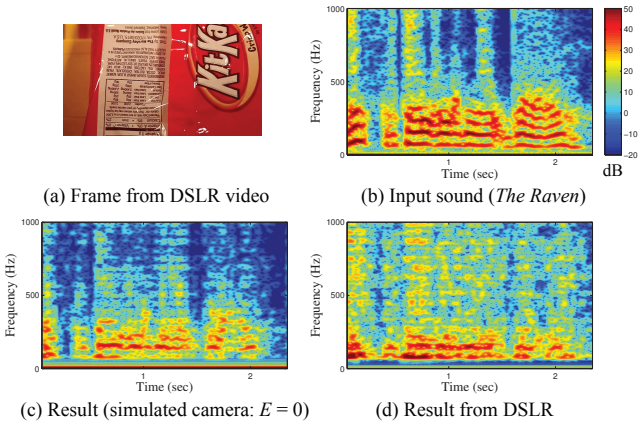
(b) Converted to audio signal

Figure 9: *Motions from a rolling shutter camera are converted to an audio signal. Each row of the video is captured at a different time. The line delay d is the time between the capture of consecutive rows. The exposure time E is the amount of time the shutter is open for each row, the frame period is the time between the start of each frame's capture and the frame delay is the time between when the last row of a frame and the first row of the next frame are captured. The motion of each row corresponds to a sample in the recovered audio signal (b). Samples that occur during the frame delay period are missing and are denoted in light gray.*

be times when the camera is not recording anything. This results in missing samples or "gaps" in the audio signal. In Fig. 9(b), we show how a triangular wave is recovered from a rolling shutter camera. Each frame contributes eleven samples, one for each row. There are five missing samples, denoted in light gray, between each frame corresponding to the nonnegligible frame delay. To deal with the missing samples in our audio signal, we use an audio interpolation technique by Janssen et al. [1986].

In practice, the exposure time is not zero and each row is the time average of its position during the exposure. For sinusoidal audio signals of frequency $\omega > \frac{1}{E}$, the recorded row will approximately be to the left of its rest position for half of the exposure and to the right for the other half. Therefore, it will not be well-characterized by a single translation, suggesting that E is a limit on the maximum frequency we can hope to capture with a rolling shutter. Most cameras have minimum exposure times on the order of 0.1 milliseconds (10 kHz).

We show an example result of sound recovered using a normal frame-rate DSLR video in Figure 10. We took a video of a bag of candy (Fig. 10(a)) near a loudspeaker playing speech, and took a video from a viewpoint orthogonal to the loudspeaker-object axis, so that the motions of the bag due to the loudspeaker would be horizontal and fronto-parallel in the camera's image plane. We used a Pentax K-01 with a 31mm lens. The camera recorded at 60 FPS at a resolution of 1280×720 with an exposure time of $\frac{1}{2000}$ seconds. By measuring the slope of a line, we determined it to have a line delay of 16 µs and a frame delay of 5 milliseconds, so that the effective sampling rate is $61920Hz$ with 30% of the samples missing. The exposure time caps the maximum recoverable frequency at around 2000Hz. In addition to audio interpolation to recover missing samples, we also denoise the signal with a speech enhancement algorithm and a lowpass filter to remove out-of-range frequencies we cannot recover due to the exposure time. We also performed a simulated experiment with identical camera parameters, except for an instant (zero) exposure time. The recovered audio clips are available online.



(a) Frame from DSLR video (b) Input sound (*The Raven*)

(c) Result (simulated camera: $E = 0$) (d) Result from DSLR

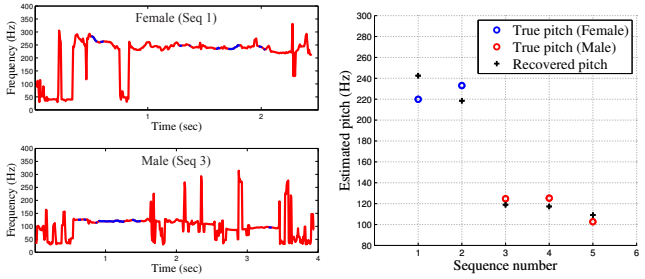
Figure 10: *Sound recovered from a normal frame-rate video, shot with a standard DSLR camera with rolling shutter. A frame from the DSLR video is shown in (a). James Earl Jones's recitation of "The Raven" by Edgar Allan Poe (spectrogram shown in (b)) is played through a loudspeaker, while an ordinary DSLR camera films a nearby Kit Kat bag. The spectrogram of the signal we manage to recover from the DSLR is shown in (d). In (c) we show the result from our rolling shutter simulation that used parameters similar to the DSLR, except for exposure time (E) that was set to zero.*

7 Discussion and Limitations

Information from Unintelligible Sound Many of our examples focus on the intelligibility of recovered sounds. However, there are situations where unintelligible sound can still be informative. For instance, identifying the number and gender of speakers in a room can be useful in some surveillance scenarios even if intelligible speech cannot be recovered. Figure 11 shows the results of an experiment where we were able to detect the gender of speakers from unintelligible speech using a standard pitch estimator [De Cheveigné and Kawahara 2002]. On our project web page we show another example where we recover music well enough for some listeners to recognize the song, though the lyrics themselves are unintelligible in the recovered sound.

Visualizing Vibration Modes Because we are recovering sound from a video, we get a spatial measurement of the audio signal at many points on the filmed object rather than a single point like a laser microphone. We can use this spatial measurement to recover the vibration modes of an object. This can be a powerful tool for structural analysis, where general deformations of an object are often expressed as superpositions of the object's vibration modes. As with sound recovery from surface vibrations, most existing techniques for recovering mode shapes are active. Stanbridge and Ewins [1999], for instance, scan a laser vibrometer in a raster pattern across a surface. Alternatively, holographic interferometry works by first recording a hologram of an object at rest, then projecting this hologram back onto the object so that surface deformations result in predictable interference patterns [Powell and Stetson 1965; Jansson et al. 1970]. Like us, Chen et al. [2014] propose recovering mode shapes from a high-speed video, but they only look at the specific case of a beam vibrating in response to being struck by a hammer.

Vibration modes are characterized by motion where all parts of an object vibrate with the same temporal frequency, the modal frequency, with a fixed phase relation between different parts of the object. We can find the modal frequencies by looking for peaks in the spectra of our local motion signals. At one of these peaks, we will have a Fourier coefficient for every spatial location in the image. These Fourier coefficients give the vibration mode shape with amplitude corresponding to the amount of motion and phase corresponding to fixed phase relation between points. In Figure 12, we map amplitude to intensity and phase to hue for two vibration modes of a drum head. These recovered vibration modes (Fig. 12(b)) closely correspond to the theoretically-derived modal shapes (Fig. 12(c)).



(a) Pitch trajectory (b) Estimated pitch

Figure 11: *Our method can be useful even when recovered speech is unintelligible. In this example, we used five TIMIT speech samples, recovered from a tissue box and a foil container. The recovered speech is difficult to understand, but using a standard pitch estimator [De Cheveigné and Kawahara 2002] we are able to recover the pitch of the speaker's voice (b). In (a) we show the estimated pitch trajectory for two recovered speech samples (female above, male below). Blue segments indicate high confidence in the estimation (see [De Cheveigné and Kawahara 2002] for details).*

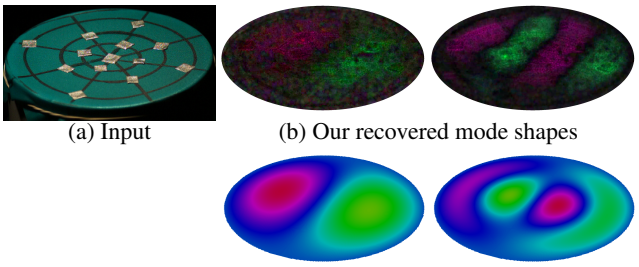
Limitations Other than sampling rate, our technique is mostly limited by the magnification of the lens. The SNR of audio recovered by our technique is proportional to the motion amplitude in pixels and the number of pixels that cover the object (Eq. 8), both of which increase as the magnification increases and decrease with object distance. As a result, to recover intelligible sound from far away objects, we may need a powerful zoom lens. The experiment in Figure 2 used a 400mm lens to recover sound from a distance of 3-4 meters. Recovery from much larger distances may require expensive optics with large focal lengths.

8 Conclusion

We have shown that the vibrations of many everyday objects in response to sound can be extracted from high speed videos and used to recover audio, turning those objects into "visual microphones". We integrate local, minute motion signals across the surface of an object to compute a single motion signal that captures vibrations of the object in response to sound over time. We then denoise this motion signal using speech enhancement and other techniques to produce a recovered audio signal. Through our experiments, we found that light and rigid objects make especially good visual microphones. We believe that using video cameras to recover and analyze sound-related vibrations in different objects will open up interesting new research and applications. Our videos, results and supplementary material are available on the project web page: `http://people.csail.mit.edu/mrub/VisualMic/`.

Acknowledgements

We thank Justin Chen for his helpful feedback, Dr. Michael Feng and Draper Laboratory for lending us their Laser Doppler Vibrometer, and the SIGGRAPH reviewers for their comments. We acknowledge funding support from QCRI and NSF CGV-1111415. Abe Davis and Neal Wadhwa were supported by the NSF Graduate Research Fellowship Program under Grant No. 1122374. Abe Davis was also supported by QCRI, and Neal Wadhwa was also supported by the MIT Department of Mathematics. Part of this work was done when Michael Rubinstein was a student at MIT, supported by the Microsoft Research PhD Fellowship.



(a) Input (b) Our recovered mode shapes

(c) Theoretically-derived mode shapes

Figure 12: *Recovered mode shapes (b) from a video of a circular latex membrane excited by a chirp playing from a nearby audio source (a). Our recovered mode shapes (b) are similar to the theoretically-derived mode shapes (c). For the modes shown in (b), the phase of surface motion across the membrane is mapped to hue, while the amplitude of vibrations across the surface is mapped to saturation and brightness.*

References

AIT-AIDER, O., BARTOLI, A., AND ANDREFF, N. 2007. Kinematics from lines in a single rolling shutter image. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, IEEE, 1–6.

BOLL, S. 1979. Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on 27*, 2, 113–120.

CHEN, J., WADHWA, N., CHA, Y.-J., DURAND, F., FREEMAN, W. T., AND BUYUKOZTURK, O. 2014. Structural modal identification through high speed camera video: Motion magnification. *Proceedings of the 32nd International Modal Analysis Conference (to appear)*.

DE CHEVEIGNÉ, A., AND KAWAHARA, H. 2002. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America 111*, 4, 1917–1930.

D'EMILIA, G., RAZZÈ, L., AND ZAPPA, E. 2013. Uncertainty analysis of high frequency image-based vibration measurements. *Measurement 46*, 8, 2630–2637.

FISHER, W. M., DODDINGTON, G. R., AND GOUDIE-MARSHALL, K. M. 1986. The darpa speech recognition research database: specifications and status. In *Proc. DARPA Workshop on speech recognition*, 93–99.

GAUTAMA, T., AND VAN HULLE, M. 2002. A phase-based approach to the estimation of the optical flow field using spatial filtering. *Neural Networks, IEEE Transactions on 13*, 5 (sep), 1127 – 1136.

GRUNDMANN, M., KWATRA, V., CASTRO, D., AND ESSA, I. 2012. Calibration-free rolling shutter removal. In *Computational Photography (ICCP), 2012 IEEE International Conference on*, IEEE, 1–8.

HANSEN, J. H., AND PELLOM, B. L. 1998. An effective quality evaluation protocol for speech enhancement algorithms. In *ICSLP*, vol. 7, 2819–2822.

JANSSEN, A., VELDHUIS, R., AND VRIES, L. 1986. Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. *Acoustics, Speech and Signal Processing, IEEE Transactions on 34*, 2, 317–330.

JANSSON, E., MOLIN, N.-E., AND SUNDIN, H. 1970. Resonances of a violin body studied by hologram interferometry and acoustical methods. *Physica scripta 2*, 6, 243.

LIU, C., TORRALBA, A., FREEMAN, W. T., DURAND, F., AND ADELSON, E. H. 2005. Motion magnification. *ACM Trans. Graph. 24* (Jul), 519–526.

LOIZOU, P. C. 2005. Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum. *Speech and Audio Processing, IEEE Transactions on 13*, 5, 857–869.

MEINGAST, M., GEYER, C., AND SASTRY, S. 2005. Geometric models of rolling-shutter cameras. *arXiv preprint cs/0503076*.

MORLIER, J., SALOM, P., AND BOS, F. 2007. New image processing tools for structural dynamic monitoring. *Key Engineering Materials 347*, 239–244.

NAKAMURA, J. 2005. *Image sensors and signal processing for digital still cameras*. CRC Press.

PORTILLA, J., AND SIMONCELLI, E. P. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vision 40*, 1 (Oct.), 49–70.

POWELL, R. L., AND STETSON, K. A. 1965. Interferometric vibration analysis by wavefront reconstruction. *JOSA 55*, 12, 1593–1597.

QUACKENBUSH, S. R., BARNWELL, T. P., AND CLEMENTS, M. A. 1988. *Objective measures of speech quality*. Prentice Hall Englewood Cliffs, NJ.

ROTHBERG, S., BAKER, J., AND HALLIWELL, N. A. 1989. Laser vibrometry: pseudo-vibrations. *Journal of Sound and Vibration 135*, 3, 516–522.

RUBINSTEIN, M. 2014. *Analysis and Visualization of Temporal Variations in Video*. PhD thesis, Massachusetts Institute of Technology.

SIMONCELLI, E. P., FREEMAN, W. T., ADELSON, E. H., AND HEEGER, D. J. 1992. Shiftable multi-scale transforms. *IEEE Trans. Info. Theory 2*, 38, 587–607.

STANBRIDGE, A., AND EWINS, D. 1999. Modal testing using a scanning laser doppler vibrometer. *Mechanical Systems and Signal Processing 13*, 2, 255–270.

TAAL, C. H., HENDRIKS, R. C., HEUSDENS, R., AND JENSEN, J. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *Audio, Speech, and Language Processing, IEEE Transactions on 19*, 7, 2125–2136.

WADHWA, N., RUBINSTEIN, M., DURAND, F., AND FREEMAN, W. T. 2013. Phase-based video motion processing. *ACM Transactions on Graphics (TOG) 32*, 4, 80.

WADHWA, N., RUBINSTEIN, M., DURAND, F., AND FREEMAN, W. T. 2014. Riesz pyramid for fast phase-based video magnification. In *Computational Photography (ICCP), 2014 IEEE International Conference on*, IEEE.

WU, H.-Y., RUBINSTEIN, M., SHIH, E., GUTTAG, J., DURAND, F., AND FREEMAN, W. 2012. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG) 31*, 4, 65.

ZALEVSKY, Z., BEIDERMAN, Y., MARGALIT, I., GINGOLD, S., TEICHER, M., MICO, V., AND GARCIA, J. 2009. Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern. *Opt. Express 17*, 24, 21566–21580.