Measuring Vibrations from Video Feeds

Problem presented by

David Smith and Andy Newell

Ministry of Defence

ESGI130 was jointly hosted by
University of Warwick
Smith Institute for Industrial Mathematics and System Engineering

Report author

Ed Brambley (University of Warwick)

Executive Summary

By using a high-speed camera, researchers at MIT in 2014 where able to recover human speech from videos of minute vibrations of objects in a room. For example, in one experiment a 2,200fps camera was positioned outside a room behind sound-proof glass, videoing an empty crisp packet on the floor inside the room, while a researcher shouted "Mary had a little lamb" at the crisp packet. By detecting minute oscillations of the crisp packet of $1\,\mu$m (0.001 mm), and using hours of computer processing, a ten second audio clip could be produced that was recognisably "Mary had a little lamb" in an American accent.

The purpose of this study group was to investigate whether this technique could be used in practice, with emphasis on the recovery of intelligible speech from a video feed of a room. During the week, the group investigated several aspects of the problem, including:

- how much an object vibrates due to sound;
- what can be done to maximize the vibration;
- how the MIT technique detects minute vibrations in videos;
- what affects the quality of the resulting recording; and
- how good a recording is needed for intelligible speech.

It was discovered the MIT experiments would not have recovered intelligible speech from an ordinary conversation; their success depended on loud sounds and prior knowledge of "Mary had a little lamb". Camera vibrations were also ignored by MIT; these are expected to be significant, but the technique could be adapted to be resilient to them. Other possibilities for enhancing their technique, by exploiting resonances or reflections, are discussed in the report. A high-speed low-noise camera is essential, and any existing video footage (such as from CCTV) is unlikely to be of sufficient quality. Further experiments with high-end high-speed cameras are needed to assess the feasibility of the technique in practice.

Version 1.0
October 9, 2017
iv+23 pages

Contributors

Ed Brambley (University of Warwick)
Helen Fletcher (University of Oxford)
Roger Hill (University of Warwick)
Ifan Johnston (University of Warwick)
Jessie Liu (University of Warwick)
Robert MacKay (University of Warwick)
James Mathews (University of Cambridge)
John Ockendon (University of Oxford)
Bernard Piette (Durham University)

Contents

1 Introduction

(1.1) In their 2014 paper, MIT researchers Davis et al. [1] demonstrated the recovery of the sound in a room from a video of some objects present in the room. The idea is that sound is the vibration of the air in the room, which causes minute vibrations of objects in the room exposed to that sound. One can then attempt to detect these vibrations from a high-speed video of the objects, and use motion-enhancement signal processing techniques developed in the same laboratory [7] to extract the audio. The aim of this report is to investigate the feasibility of using this technique to extract intelligible speech in practical situations.

(1.2) Section 2 of the report analyses the vibration of simple objects, such as the ones used by Davis et al. [1], in order to model the amplitude of vibration as a function of the size and material properties of the object and the amplitude and frequencies of the sound. In addition to giving ballpark figures of what sort of equipment would be needed to detect the vibrations, one other aim of this modelling is to determine the ideal properties that an object would have in order to be used as a *visual microphone*. One novel possibility is to gain greater sensitivity to motion by looking at reflections in an object, rather than the object itself; this is considered further in section 3.

(1.3) Section 4 describes the earlier work from the MIT lab on which the recovery of sound is based. Wadhwa et al. [7] developed a technique to analyse video and enhance the motion shown in the video in a particular frequency range. This is how Davis et al. [1] were able to detect the tiny motion of objects due to sound.

(1.4) Section 5 investigates what is required for a recording of speech to be intelligible. This also gives ballpark figures on what frequencies and noise levels are needed in practice.

(1.5) Finally, section 6 summarizes the results in this report, and suggests further lines of inquiry.

1.1 Sound during conversations

(1.6) The human voice consists of frequencies ranging from 80 Hz to 4 kHz excluding sibilants. In telephony, the voice band is approximately 300 Hz to 3.4 kHz[1], with the missing information below 300 Hz perceived as a missing fundamental[2]. Sound restricted to the voice band is noticeably telephone-like, but is none-the-less fully intelligible.

[1] `http://en.wikipedia.org/wiki/Voice_frequency`

[2] `http://en.wikipedia.org/wiki/Missing_fundamental`

0 dB	Threshold of human hearing
40 dB	Library background noise
50 dB	Quiet conversation
60 dB	Conversation against background noise
70 dB	Vacuum cleaner
80 dB	Freight train at 15 metres
110 dB	Jet aircraft at 100 metres, and the threshold of pain

Table 1: Sound volume on a deciBel (dB) scale. Excerpts taken from `http://www.industrialnoisecontrol.com/comparative-noise-examples.htm`.

(1.7) Sound in air is a wave, consisting of small oscillating motion of the air particles and a corresponding small oscillation of air pressure. Sound volume is measured using the logarithmic deciBel (dB) scale, shown in table 1. The corresponding maximum displacement of an air particle, ξ, due to the sound is approximately given by

$$\xi = \frac{\sqrt{2}}{\pi f \rho_0 c_0} 10^{\frac{\mathrm{dB}}{20}-5} \approx \frac{1}{f} 10^{\frac{\mathrm{dB}}{20}-8} \tag{1}$$

where f is the frequency of the sound in Hertz, dB is the sound amplitude in deciBels, ρ_0 and c_0 are the density ($1.2\,\mathrm{kg/m^3}$) and sound speed ($340\,\mathrm{m/s}$) of air, and ξ is given in metres.

(1.8) As an example of the minute motion of objects due to sound, a loud conversations at 60 dB at a typical frequency of 300 Hz would cause the air to move by approximately 30 nm, or 1/2000th of a human hair.

2 Acoustic excitations of thin plates

(2.1) In this section, we investigate models of sound interacting with an object. The aim is to predict the amplitude of motion of the object given the incident sound, and hence to predict parameters that would make for a good visual microphone. First, in section 2.1, we consider sound interacting with an infinite thin plate. This allows us to investigate how much of the sound reflects back from the plate, and how much the plate moves, for plates of different materials. Since the infinite plate model ignores resonances, which could significantly increase the plate motion, in section 2.2 we investigate resonances of a rectangular section of plate simply supported at its edges. How this plate would react to forcing is considered in section 2.3, although this over estimates the amplitudes of the plate at resonance since it does not take account of the back reaction of the plate's motion on the air. In section 2.4, we consider adding artificial spring and damping terms to the infinite plate model of section 2.1 in order to assess the interactions between the sound and the plate resonances, albeit using a rather artificial model.

Material	density (kg/m^3)	E (GPa)	ν
Polyethylene (Low Density)	950	0.95	0.4
Polystyrene foam	25-45	1.9-2.9	0.4
Silica aerogel	1	10	0.33
Aluminium foil	2300	70	0.33
rubber	1522	0.01-0.1	0.48 - 0.5
Glass	2400-2800	50-90	0.2 - 0.27
Aluminium	2700	69	0.334
Copper	8790	117	0.355
Steel	7820	180	0.265-0.305
Tin	7280	47	0.33
MDF	700-720	4	0.25
Pine wood	554-740	9	0.3-0.4

Table 2: Approximate elastic properties of some relevant materials.

Ideally, one would add the back reaction of the air into the finite plate model of section 2.3, although this is sufficiently complicated that it was beyond the scope of the one-week study group.

(2.2) For reference in what follows, the equation for the displacement u of a vibrating plate is given by Landau and Lifshitz [3] as

$$h\rho\frac{\partial^2 u}{\partial t^2} + B\nabla^4 u = P \tag{2}$$

where ρ is the plate density (kg/m^3), h is the plate thickness in metres, P is the net force per unit area acting on the plate in Pascals, and

$$B = \frac{Eh^3}{12(1-\nu)} \tag{3}$$

is the bending stiffness, where E is the Young's modulus of the plate material and ν its Poisson ratio. Some approximate physical properties of relevant materials are given in table 2.

2.1 Sound exciting an infinite elastic beam

(2.3) In this section, we simplify the object to an infinite vertical elastic beam. We assume that a plane wave of frequency ω is incident from the left with an incident angle θ, as shown in figure 1. Part of the incident wave is reflected back to the left, and part of it is transmitted. The pressure of the incoming wave is of the form

$$P_{\text{inc}}(x,y,t) = \text{Re}\Big(P_0 \exp\left\{\text{i}(\omega t - k_x x - k_y y)\right\}\Big) \tag{4}$$

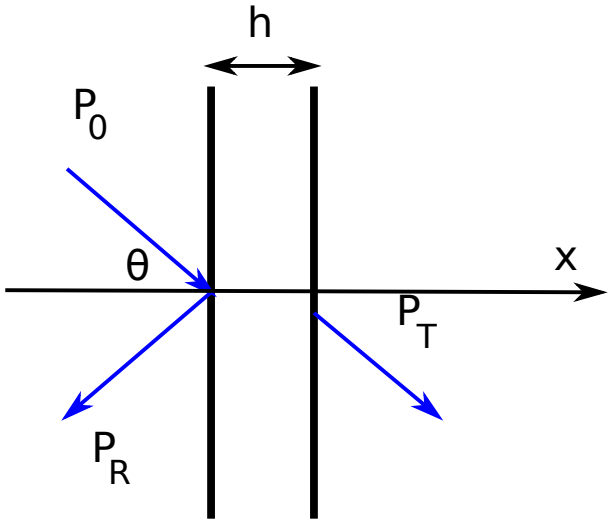

Figure 1: Scattering of a plane wave on a thin beam of thickness h. Incident and reflected wave on the left, transmitted wave on the right.

where

$$P_0 = 2\sqrt{2} \times 10^{-5} \times 10^{\mathrm{dB}/20}, \qquad k_x = \frac{\omega}{c_0}\cos\theta, \qquad k_y = \frac{\omega}{c_0}\sin\theta. \tag{5}$$

The reflected and transmitted waves are therefore of the form

$$\begin{aligned} P_{\mathrm{ref}}(x,y,t) &= \mathrm{Re}\Big(P_R \exp\left\{\mathrm{i}(\omega t + k_x x - k_y y)\right\}\Big), \\ P_{\mathrm{trans}}(x,y,t) &= \mathrm{Re}\Big(P_T \exp\left\{\mathrm{i}(\omega t - k_x x - k_y y)\right\}\Big), \end{aligned} \tag{6}$$

with the corresponding horizontal velocities given by

$$\begin{aligned} v_{\mathrm{inc}}(x,y,t) &= \mathrm{Re}\left(\frac{P_0\cos\theta}{\rho_0 c_0}\exp\left\{\mathrm{i}(\omega t - k_x x - k_y y)\right\}\right) \\ v_{\mathrm{ref}}(x,y,t) &= \mathrm{Re}\left(-\frac{P_R\cos\theta}{\rho_0 c_0}\exp\left\{\mathrm{i}(\omega t + k_x x - k_y y)\right\}\right) \\ v_{\mathrm{trans}}(x,y,t) &= \mathrm{Re}\left(\frac{P_T\cos\theta}{\rho_0 c_0}\exp\left\{\mathrm{i}(\omega t - k_x x - k_y y)\right\}\right) \end{aligned} \tag{7}$$

where ρ_0 and c_0 are respectively the density and the sound speed of air. This notation uses complex amplitudes P_0, P_T and P_R to describe both the amplitude and phase of the solutions for convenience, despite the underlying quantities for pressure and velocity being real.

(2.4) We call the displacement of the beam $u(y,t)$ and we then impose that the displacement of the air and the beam must match on both sides of the beam

$$v_{\mathrm{inc}}(0,y,t) + v_{\mathrm{ref}}(0,y,t) = \frac{\partial u(y,t)}{\partial t} = v_{\mathrm{trans}}(0,y,t) \tag{8}$$

This implies that $u(y,t) = \mathrm{Re}\Big(U \exp\big\{\mathrm{i}(\omega t - k_y y)\big\}\Big)$, with

$$(P_0 - P_R)\cos\theta = \mathrm{i}\omega\rho_0 c_0 U = P_T \cos\theta \tag{9}$$

Equation (2) gives Newton's law of motion applied to the beam,

$$h\rho\frac{\partial^2 u}{\partial t^2} + B\frac{\partial^4 u}{\partial y^4} = P_{\text{inc}} + P_{\text{ref}} - P_{\text{trans}}. \tag{10}$$

Balancing the forces on the beam using (10) means that U must also satisfy

$$\big(-\omega^2 h\rho + Bk_y^4\big)U = P_0 + P_R - P_T. \tag{11}$$

Solving (9) and (11) simultaneously leads to

$$P_T = \frac{\mathrm{i}\rho_0 c_0 \omega}{\cos\theta}U \qquad P_R = P_0 - \frac{\mathrm{i}\rho_0 c_0\omega}{\cos\theta}U \tag{12}$$

$$U = \frac{2P_0}{\frac{B\omega^4}{c_0^4}\sin^4\theta - \omega^2 h\rho + \mathrm{i}\omega\frac{2\rho_0 c_0}{\cos\theta}}. \tag{13}$$

Equation (13) therefore gives the amplitude and phase of the oscillation of the beam, subjected to an incoming wave of amplitude P_0. Ideally, therefore, we would like the amplitude $|U|$ to be as large as possible to be most easily detected.

(2.5) Since (13) is relatively complicated, it is helpful to look at some simplifying cases. In particular, for a wave perpendicular to the beam ($\sin\theta = 0$) the bending stiffness of the beam is unimportant, and we have

$$U = \frac{2P_0}{-\omega^2 h\rho + 2\mathrm{i}\omega\rho_0 c_0}. \tag{14}$$

In the limit of a very thin beam, or a very light beam, then $h\rho \to 0$, and we recover $|U| = |P_0|/(\omega\rho_0 c_0)$, which is the expression for the displacement of the air; that is, a light beam moves with the air. For a heavier or thicker beam, the motion is smaller, especially at higher frequencies. The order of magnitude of the frequency when the beam stops moving with the air is given by $f_c = 2\pi\omega_c \sim 4\pi\rho_0 c_0/(h\rho)$. For frequencies much lower than this critical frequency f_c the beam moves with the air, while for frequencies much higher than f_c the beam moves much less than the air.

(2.6) Figure 2 plots the amplitude of oscillation $|U|$ of various beams. The first two sub-figures are for 50 μm thick polyethylene, emulating a crisp packet. Louder amplitudes of sound lead to larger displacements, and higher frequencies lead to smaller displacements; this is expected, as a sound wave in air has the same displacement profile. The displacement is relatively insensitive to the direction of the incident sound, provided it is not parallel to the surface ($\theta = 90°$). The final sub-figure in figure 2 shows that several materials

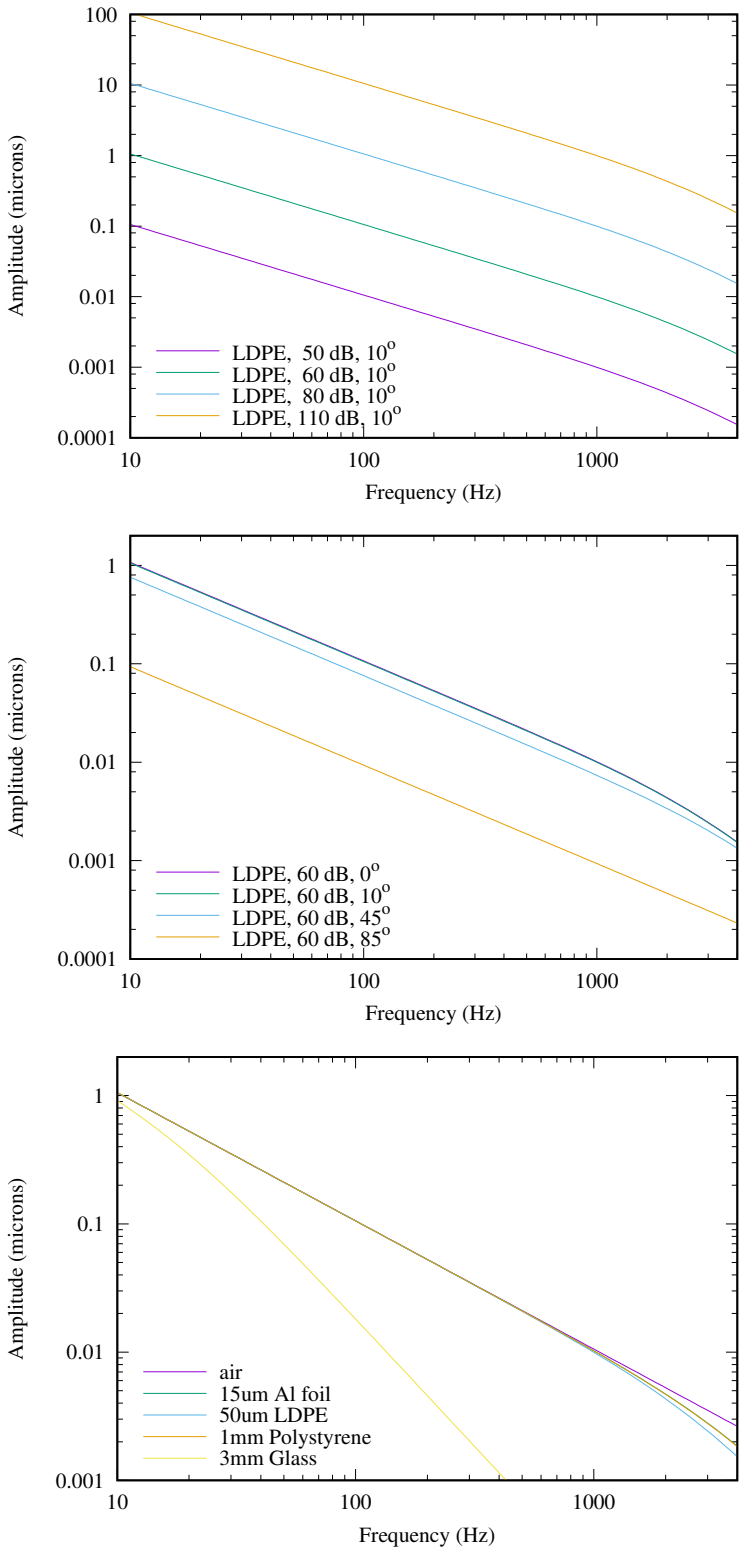

Figure 2: Displacement of an acoustic wave bouncing off a thin beam. Top: a crisp packet (50 μm thick polyethylene) at various amplitudes. Middle: a crisp packet subjected to 60 dB sound at various angles. Bottom: various materials subjected to 60 dB sound at 10° angle of incidence.

act the same, and effectively just act as passive tracers of the vibration in the air, at least at low to moderate frequencies, while high frequencies are more attenuated. This is not the case for all materials, however; 3 mm thick glass has a much smaller amplitude, especially at higher frequencies.

(2.7) This model is of course a rather simple one. For example, it assumes that the object is an infinite flat beam with no curvature or edges, that the beam is not clamped or pinned in any way, and that the incoming wave is only present to the left of the beam. Some of these could be incorporated as extensions of this model. This model also ignores resonances as is assumes an infinite beam. Resonant frequencies occur in all finite elastic scatterers, causing the acoustic displacement to be much greater than at other non-resonant frequencies. In order to investigate resonance, we next describe a model of a finite size elastic plate and its resonances.

2.2 Resonances of a 2D rectangular plate

(2.8) In this section, we develop a model of the resonances of an elastic plate supported by its edges. We assume a rectangular plate of size $L_x \times L_y$ resting on its edges and solve the unforced 2D version of (2),

$$h\rho\frac{\partial^2 u}{\partial t^2} + B\nabla^4 u = 0. \tag{15}$$

We use separation of variables and take $u(t,x,y) = g(t)W(x,y)$. Then

$$\frac{h\rho}{g}\frac{\partial^2 g}{\partial t^2} = -\frac{B}{W}\left(\frac{\partial^4 W}{\partial x^4} + \frac{\partial^4 W}{\partial y^4} + 2\frac{\partial^4 W}{\partial x^2 \partial y^2}\right). \tag{16}$$

For equality to hold, both sides must be a constant, say $-\omega^2$, and we may therefore take $g(t) = \sin(\omega t)$. Then

$$\frac{B}{\rho h}\left(\frac{\partial^4 W}{\partial x^4} + \frac{\partial^4 W}{\partial y^4} + 2\frac{\partial^4 W}{\partial x^2 \partial y^2}\right) = \omega^2 W. \tag{17}$$

We then impose the boundary conditions for a freely supported resting plate with no bending moments at the edges:

$$W(0,y) = W(L_x,y) = W(x,0) = W(x,L_y) = 0 \tag{18}$$

$$B\left(\frac{\partial^2 W}{\partial x^2} + \nu\frac{\partial^2 W}{\partial y^2}\right) = 0, \qquad \text{at } x = 0 \text{ and } x = L_x \tag{19}$$

$$B\left(\frac{\partial^2 W}{\partial y^2} + \nu\frac{\partial^2 W}{\partial x^2}\right) = 0, \qquad \text{at } y = 0 \text{ and } y = L_x. \tag{20}$$

We then notice that $W(x,y) = A\sin(k_x x)\sin(k_y y)$ satisfies trivially all the above boundary conditions if $k_x = n\pi/L_x$ and $k_y = m\pi/L_y$ where n and m

are positive integers. Substituting this ansatz into (16), we get

$$\omega = \pi^2 \left(\frac{n^2}{L_x^2} + \frac{m^2}{L_y^2} \right) \sqrt{\frac{B}{\rho h}}. \tag{21}$$

For a polyethylene plate with $L_x = L_y = 0.1\,\mathrm{m}$ and $h = 50\,\mu\mathrm{m}$, we have $B \approx 1.65 \times 10^{-5}$, and this gives resonant frequencies $f = \omega/(2\pi)$ of

$$f \approx 2.93(n^2 + m^2)\,\mathrm{Hz}. \tag{22}$$

Resonant frequencies may therefore be expected to be rather common, and hence the acoustic displacement of an object may well be much closer to the acoustic displacement of the air than might otherwise have been thought without considering resonances.

(2.9) It should be noted that this model assumes the plate is flat, is supported only at its edges, and that there is no friction or loss at the edges. Again, such extensions could be incorporated into a more complicated model, but it is unlikely that the exact resonant frequencies will be of use in practice; rather, if the resonant frequencies were to be used explicitly in the algorithm for extracting sound from image motion, one would need to best-fit the resonant frequencies given the response of the object when forced by the sound in the room. The forcing of this resonant plate is considered in the next section.

2.3 A forced elastic plate

(2.10) We now consider the elastic plate from the previous section subjected to an external forcing. For simplicity, in this section we consider only the 1D problem, so that the governing equation is

$$h\rho \frac{\partial^2 u}{\partial t^2} + B \frac{\partial^4 u}{\partial x^4} = P(x,t), \tag{23}$$

where $P(x,t)$ is the force per unit area. Since the problem is linear, we may without loss of generality assume the wave has a single frequency, $P(x,t) = P(x)\sin(\omega t)$, since multiple frequencies may be summed over if required. If the plate has length L, the eigenmodes of a resting plate are given by (21) as $u(x,t) = A\sin(\omega_j t)\sin(k_j x)$, with

$$k_j = \frac{j\pi}{L}, \qquad \omega_j = \frac{j^2\pi^2}{L^2}\sqrt{\frac{B}{h\rho}}. \tag{24}$$

The solutions of (23) can be written as

$$u = \sum_{j=1}^{\infty} \frac{A_j \sin(\frac{j\pi x}{L})\sin(\omega t)}{h\rho(\omega_j^2 - \omega^2)}, \tag{25}$$

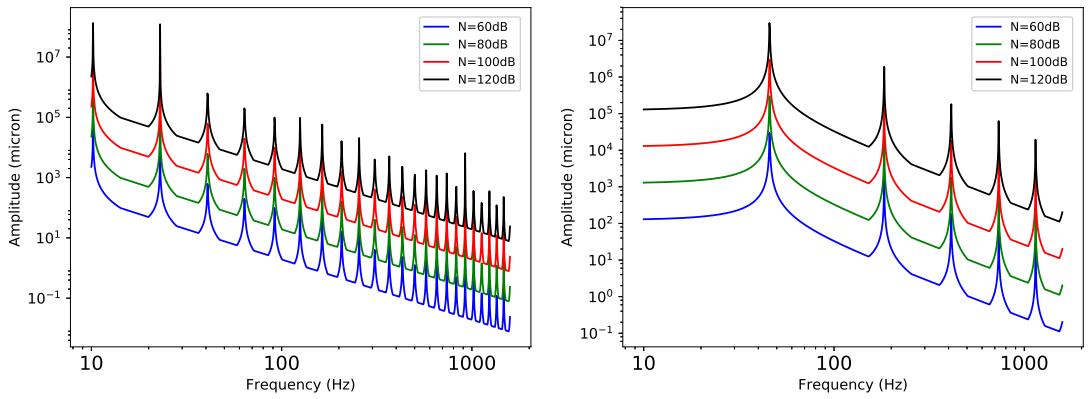

Figure 3: Maximum plate amplitude after excitation as a function of incoming frequency for a 10 cm long thin plate of a) 50 μm polyethylene, and b) 100 μm polystyrene foam.

which may be seen by substituting (25) into (23) to get

$$\begin{aligned} P(x) &= \sum_{j=1}^{\infty} A_j \left(\frac{-h\rho\omega^2 + B\left(\frac{j\pi}{L}\right)^4}{h\rho\omega_j^2 - h\rho\omega^2} \right) \sin\left(\frac{j\pi x}{L}\right) \\ &= \sum_{j=1}^{\infty} A_j \sin\left(\frac{j\pi x}{L}\right). \end{aligned} \tag{26}$$

The A_j coefficients are therefore the Fourier series coefficients of the function $P(x)$,

$$A_j = \frac{2}{L}\int_0^L \sin\left(\frac{j\pi x}{L}\right) P(x)\,\mathrm{d}x. \tag{27}$$

In particular, if the wavelengths of the sound in the air are much longer than the length L, then the pressure $P(x)$ may be taken as a constant, $P(x) = P_0$, and then

$$A_j = \begin{cases} \frac{4P_0}{\pi j} & \text{for odd } j \\ 0 & \text{for even } j \end{cases}. \tag{28}$$

Note that, if the frequency of excitation ω is the same as one of the resonance of the plate ω_j, then equation (25) predicts an infinite amplitude of oscillation of the plate. This is because the radiation damping from the back reaction of the plate movement on the forcing has not yet been included.

(2.11) Figure 3 plots some examples of the forced response of plates of various materials. Unlike figure 3, different points on the plate are moving with different amplitudes, and so figure 3 plots the maximum amplitude occurring anywhere on the plate. The resonant frequencies are clearly visible as the peaks in figure 3. Most importantly, the number of resonant peaks is seen to be very important; polystyrene foam is lighter than polyethylene, but when resonances are included polyethylene oscillates more than than polystyrene. This is perhaps why the MIT researchers [1] recovered better results from a

crisp packet (made of polyethylene) than they did from a disposable drinks cup (made of polystyrene foam). Note that the amplitudes of oscillation in figure 3 are larger than those in figure 2, since in figure 2 energy is lost by radiating sound back into the air.

(2.12) In this section, we have neglected any dissipation that might limit resonance, such as dissipation within the air or friction of the plate with its supports. The forcing was also considered given and the response of the plate was calculated; this neglects the back reaction of the plate movement on the wave in the air, which will also limit the amplitude at resonance. This is addressed in the next section.

2.4 Interaction between sound and a resonating object

(2.13) In the analysis above, section 2.1 accounts for the back reaction of the plate on the air (through wave reflection and transmission), but ignores resonances. Contrastingly, section 2.3 includes resonances, but ignores the back reaction of the plate on the air, leading to arbitrarily large plate motion at the resonant frequencies. In this section, we modify the model in section 2.1 to include an artificial spring and damping term, in order to investigate the combination of back reaction and resonance.

(2.14) We modify Newton's law for the beam (10) to include an artificial spring term $h\rho\omega_0^2$ (giving an undamped resonance at frequency ω_0) and an artificial damping μ. One could think of this as a crude model of a horizontal beam lying on a carpet, with the carpet providing the extra spring and damping terms, and the transmitted wave being totally absorbed by the carpet without reflection. The resulting governing equation is

$$h\rho\frac{\partial^2 u}{\partial t^2} + \mu\frac{\partial u}{\partial t} + h\rho\omega_0^2 u + B\frac{\partial^4 u}{\partial y^4} = P_{\text{inc}} + P_{\text{ref}} - P_{\text{trans}}. \tag{29}$$

By following the same method as in section 2.1, we arrive at the equivalent of equation (13),

$$U = \frac{2P_0}{\frac{B\omega^4}{c_0^4}\sin^4\theta + (\omega_0^2 - \omega^2)h\rho + \mathrm{i}\omega\left(\mu + \frac{2\rho_0 c_0}{\cos\theta}\right)}, \tag{30}$$

or, for a wave perpendicular to the beam ($\sin\theta = 0$), the equivalent of equation (14),

$$U = \frac{2P_0}{(\omega_0^2 - \omega^2)h\rho + \mathrm{i}\omega(\mu + 2\rho_0 c_0)}. \tag{31}$$

This shows that the $2\mathrm{i}\omega\rho_0 c_0$ term found previously is a radiation damping term, which is increased by adding the artificial damping μ, while the resonance at ω_0 can cancel out the mass of the beam so as to give results near resonance as if the beam were much lighter. Without artificial damping

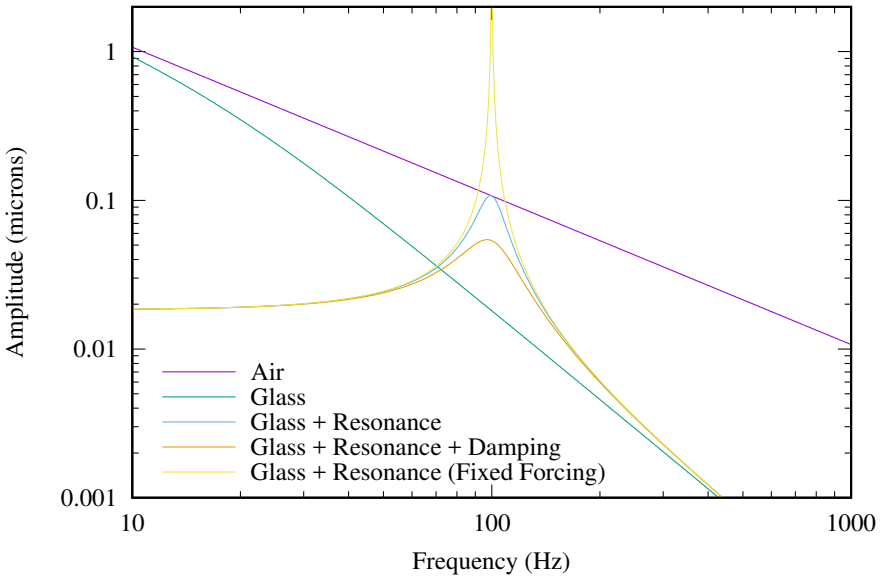

Figure 4: Amplitude of oscillations of a beam with an artificial spring (resonant frequency ω_0) and damping (strength μ), using (31). Air is the displacement of the air for the incoming wave. Glass is for 3 mm thick glass. Glass+Resonance is for glass with an artificial spring resonating at 100 Hz. Glass+Resonance+Damping adds an additional damping of strength $\mu = 2\rho_0 c_0$ (comparable to the radiation damping). "Fixed forcing" assumes no back reaction of the beam on the air (as in section 2.3).

($\mu = 0$), at resonance we recover $U = P_0/(\mathrm{i}\omega\rho_0 c_0)$, which is the displacement amplitude of the air. Without both artificial and radiation damping (setting $\mu = \rho_0 = 0$), we find an infinite beam amplitude at resonance when $\omega = \omega_0$, in agreement with the results of the previous section.

(2.15) Figure 4 plots the results of this for a 3 mm thick glass beam. With an artificial resonance at 100 Hz, the amplitude of oscillation of the beam at 100 Hz is the same as that of the air, while without the artificial resonance the beam amplitude is ten times smaller. Adding extra dissipation reduces the amplitude at resonance, while ignoring the radiation damping (by using a fixed forcing as in section 2.3) gives the expected infinite amplitude at resonance.

(2.16) Importantly, note that introducing a resonance can also have negative effects, such as anti-resonance. This can be seen at low frequencies in figure 4, where the amplitude of the glass with artificial resonance is much smaller than the amplitude of the glass without artificial resonance. As is known from current research on generating electricity from water waves, designing resonant systems to capture energy from waves is far from easy.

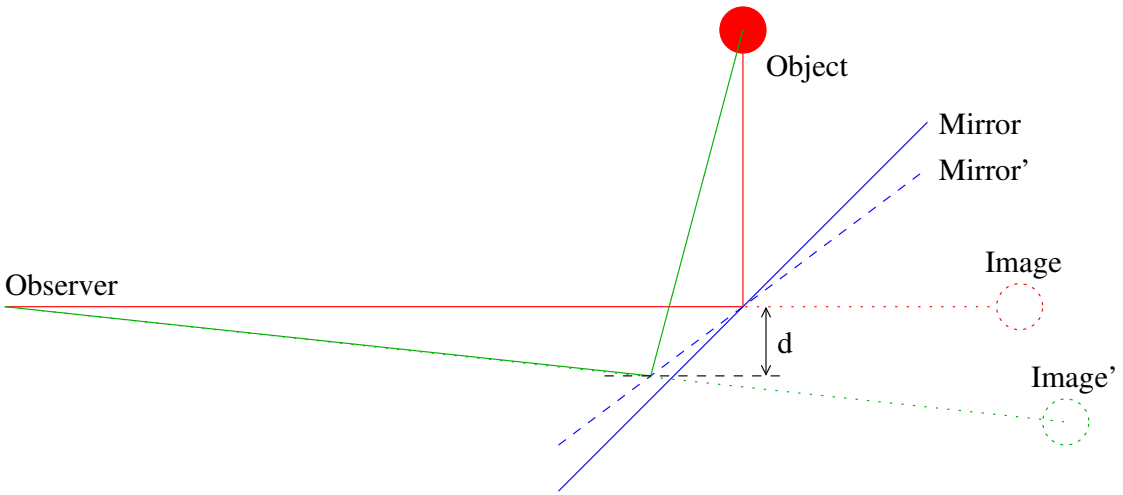

Figure 5: An observer sees an object (O) in a mirror (M) as if it were the image (I) behind the mirror. When the mirror moves to the primed position (M'), the image moves, and the magnitude of the motion is the same as a motion at the mirror of magnitude d.

3 Reflection from a bending mirror

(3.1) It may be that looking at the motion of reflections allows for greater sensitivity than simply looking for lateral motion. In this section, we investigate this by considering the motion of images in an oscillating mirror. Figure 5 shows an object (O) seen by an observer in a mirror (M)[3]. Using complex numbers for 2D coordinates, if a point on the mirror is at location $m = m_x + \mathrm{i}m_y$ and at an angle θ to the vertical, and the object is located at $z = x + \mathrm{i}y$, then the image (I) of the object seen by the observer at the origin is given by

$$I = m + \overline{(z-m)}\mathrm{e}^{-2\mathrm{i}\theta}, \tag{32}$$

where an overbar denotes the complex conjugate. If the mirror then rotates to a second angle θ', the distance the image appears to move in the mirror, d, is given by

$$d = |m|\tan(\arg q) = |m|\frac{\operatorname{Im} q}{\operatorname{Re} q} \qquad q = \frac{m + \overline{(z-m)}\mathrm{e}^{-2\mathrm{i}\theta'}}{m + \overline{(z-m)}\mathrm{e}^{-2\mathrm{i}\theta}}. \tag{33}$$

It is helpful to rearrange this in terms of the original image position I,

$$\overline{z-m} = (I-m)\mathrm{e}^{2\mathrm{i}\theta} \qquad \Rightarrow \qquad q = \frac{m}{I} + \left(1 - \frac{m}{I}\right)\mathrm{e}^{2\mathrm{i}(\theta-\theta')}. \tag{34}$$

[3]Even though the elastic displacements (of the order of 1 μm) are comparable to the wavelengths of visible light (of the order of 0.4 μm), so that using ray theory for the light from the mirror would be inappropriate, figure 5 may be thought of as using the method of images to solve for a perfectly reflecting boundary condition on the mirror, which remains valid for light of any wavelength.

m/I is the distance to the mirror normalized by the distance to the image. Since the image is always "behind" the mirror, this ratio always has modulus less than one. If the image is a significant distance away, such as the reflection of the sun, moon, or clouds, then $|m/I| \ll 1$, and in this case for small angular changes $\theta - \theta'$, we find $|d| = 2|m||\theta - \theta'|$. This is to be expected; if you were looking at yourself in a hand mirror, and then turned the hand mirror 45° upwards, you would see the ceiling in the mirror, which is a $90° = 2 \times 45°$ change in direction.

(3.2) How does this compare with the motion of an actual object? For example, are we better to look at the motion of a crisp packet, or the motion of reflections in the crisp packet? In order to answer this, we consider a mirrored bending beam. The bottom of the beam is fixed, while the top of the beam has moved a distance a. The shape of the beam is therefore given by $x = ay^2/\ell^2$, where ℓ is the length of the beam. The angle of the top of the beam, for small deflections, is approximately $\theta' \approx \frac{\mathrm{d}x}{\mathrm{d}y}|_{y=\ell} = 2a/\ell$, and therefore the displacement of a reflection in the top of the beam is $d \approx 4|m|a/\ell$; that is, the displacement a is magnified by a factor $4|m|/\ell$ when looking at the reflection, where ℓ is the length of the beam (e.g. the size of the object) and $|m|$ is the distance to the camera. Clearly $|m| \gg \ell$, and hence tracking moving reflections in objects is predicted to lead to significantly better sensitivity than just tracking lateral motion of the object. As an example, for an object of size $\ell = 10\,\mathrm{cm}$ oscillating with $1\,\mu\mathrm{m}$ amplitude viewed from a camera $|m| = 10\,\mathrm{m}$ away, the effective motion of reflections in the object is predicted to be of the order $d = 400\,\mu\mathrm{m}$, which should easily be detectable.

(3.3) This section assumes that there are suitable objects in a room to cause reflections (such as lights, windows, etc), and that motion of reflections may be detected as easily as motion of the object itself. This latter assumption is probably quite limiting, since diffusive reflections may be much harder to get accurate motion from. Practical tests with oscillating reflective objects would be helpful to test the validity of these assumptions.

4 Detecting motion from video

(4.1) The underlying process behind the visual microphone MIT paper [1] relies on being able to amplify the motion of object in a video at certain frequencies; this is described by a previous publication by MIT researchers Wadhwa et al. [7]. The process makes use of a "complex-valued steerable pyramid" wavelet decomposition [4–6]. As described by [5], the is overcomplete (i.e. produces more bytes of data than the input), but is invertible (so that the original image can be reconstructed from the wavelet coefficients). The signal is separated into high-, low-, and band-limited spatial frequencies. The high-frequencies are stored unencoded. The band-limited frequencies are

encoded using several different positions and orientations of wavelets. The low-frequencies are downsampled to half the resolution, and the process repeated (hence the pyramid structure). This ensures details of the image at several different scales and at several different orientations is produces.

(4.2) Just as for a Fourier series, Wadhwa et al. [7] claim that the complex coefficients of the resulting transform can be separated into their amplitude and phase, with a change in phase corresponding to translation. By transforming each frame of a video, the phase of each coefficient can have particular temporal frequencies amplified, which then amplifies the motion in the image at these frequencies. The same technique of using the phase of each coefficient was used for the visual microphone paper by Davis et al. [1].

(4.3) Because of the pyramid structure of the wavelets, information about the motion of the entire image will be encoded using the coefficients at the bottom of the pyramid. While these coefficients were not treated differently than the other coefficients in the MIT papers, it is likely that using these coefficients carefully could eliminate camera movements from the signal; although this was not investigated further here.

(4.4) To investigate the detection of motion from videos, we take the code from Ref. 7, available online[4], and investigate some of the results from the paper, and our own example. For our own example, we excite a projector screen of approximate height $h = 3.5$m, and film the oscillations with both a DSLR camera on a tripod and a hand-held mobile phone camera. The frame rate for both was 30fps, with a video size of 960×480 pixels. We take advantage of the first few resonant frequencies of the screen. Treating the screen as a simple pendulum gives an (angular) frequency of $\sqrt{h/g}$, while adding in some effects of torsion instead gives $\sqrt{3h/g}$ as the angular frequency. Converting to Hz then gives frequencies in the range 0.26-0.46Hz. We therefore choose to selectively amplify from 0.2Hz to 0.6Hz when using the MIT software. The running time on a standard laptop were on the order of several minutes, depending on the video length and number of frames; our processing used a reduced resolution video to speed up the computation, as can be seen in the figures below when comparing the original and motion enhanced images.

(4.5) We get good results for both the mobile phone footage and DSLR footage, provided the cameras are held steady. We recover the oscillations of the projector at the expected frequencies. Some stills from these movies are displayed in Figure 6.

(4.6) We are able to detect and magnify the motion, even when the video frame rate is reduced to 2fps. We do this by just selected the 15-th and 30-th frame per second.

[4] `http://people.csail.mit.edu/nwadhwa/phase-video/PhaseBasedRelease_20131023.zip`

(a) Original image (b) Motion magnified image

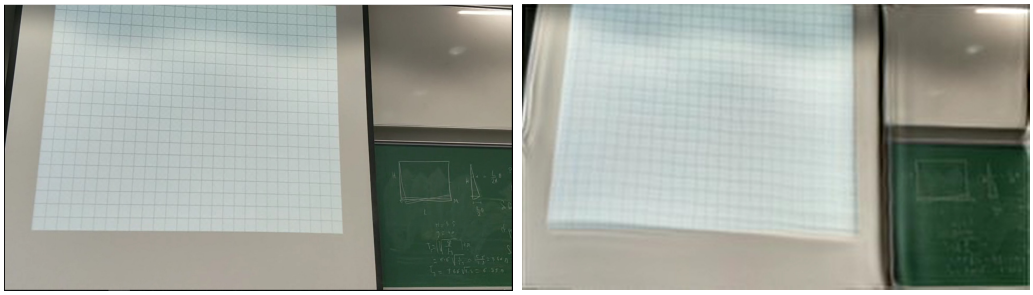

Figure 6: Stills from the unmagnified (left) and motion magnified (right) videos of a moving projector screen. Note that, to save processing time, the quality of the video on the right has been reduced.

(4.7) We are able to add noise to the original video of a crane that appears to be stationary, and still detect the motion of the crane at 0.2Hz to 0.4Hz as discussed in Wadhwa et al. [7]. It was pointed out in Wadhwa et al. [7] that this is entirely expected, since the technique may redistribute noise but will never amplify it. The noise we added was Gaussian white noise with zero mean and variance $\sigma^2 = 0.01$, using the "imnoise" command in Matlab.

(4.8) If the camera is moving (such as a hand held mobile phone footage), we are unable to get sensible results due to the motion of the camera. We see the whole image moving, and it is difficult to detect what is still and what is not after applying the algorithm. Pre processing to reduce the movement of the camera might improve the algorithm, but it was not tested here.

(4.9) We conclude that the software is working reasonably well and able to be used on a standard laptop. Using larger image sizes requires more memory, which was the main limiting factor in the image size we choose.

4.1 Rolling shutter

(4.10) We now investigate the effect of a rolling shutter, which allows the recovery of higher frequencies that the frame rate. This technique was suggested in Davis et al. [1].

(4.11) For example, let us consider the case of a 50fps standard camera, and wanted to detect motion at 200Hz, i.e four times faster than the frame rate. For simplicity, we assume we have a couple of objects which move from position 1 to position 2 at a frequency of 200Hz.

(4.12) If we have a global shutter, then we capture all the information in the image at one moment. Thus, we would only see the image at position 1 if we capture the image using the global shutter.

(4.13) If we instead have a rolling shutter, then each line is exposed for a small amount of time (with a typical minimum exposure time of 1/2000s, so in

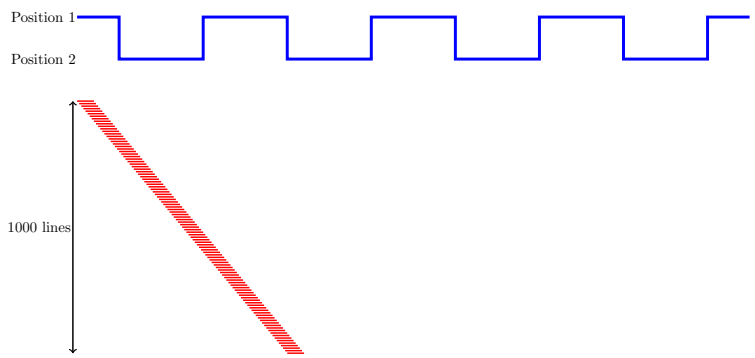

Figure 7: Schematic of rolling shutter for a single frame.

(a) Original image (b) Magnifying image

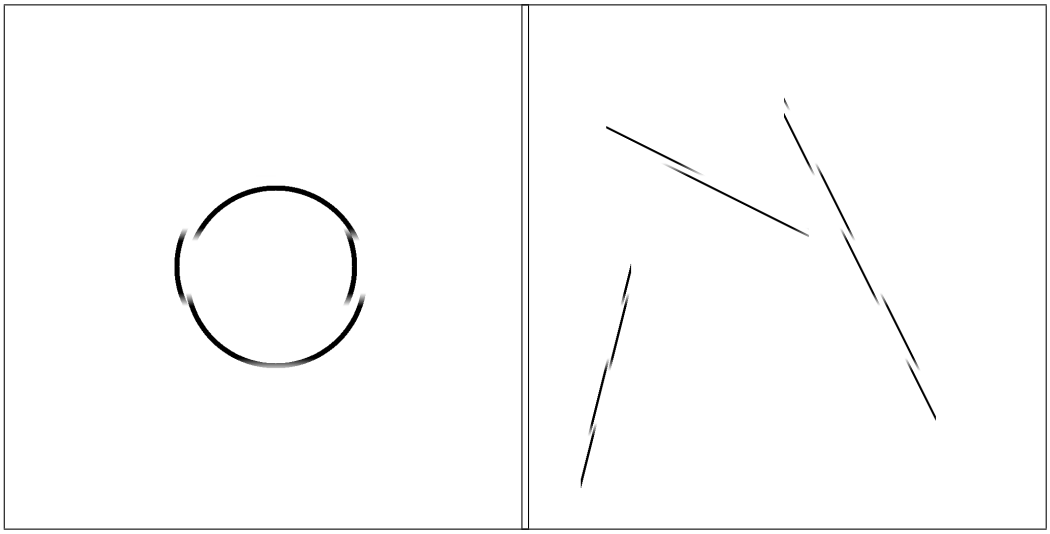

Figure 8: Results from using a rolling shutter to detect the frequency of objects moving from position 1 to position 2.

this case 1/40th of a frame). Each subsequent line is then offset by a frame delay, which would depend on the number of lines of the camera. Figure 7 illustrates the motion and rolling shutter for a single frequency.

(4.14) The result of using the rolling shutter are displayed in Figure 8, for two toy examples of a circle and multiple lines moving from position 1 to position 2. We can clearly see the effect of the rolling shutter. When the object changes position during the exposure time of a single line, we take the average value which is the result of the grey parts of the image.

(4.15) The problem is then to recover the frequency signal in Figure 7 from the pictures in Figure 8. From Figure 8a we can clearly see this is difficult, as the image is only in the centre of frame, while it is more possible in Figure 8b. Clearly, there is also more than one frequency and original images we can deduce, so in that sense we have an inverse problem. For more realistic cases, we would expect the frequency signal between two positions to be a

sine wave rather than a square signal, which would result in a curved line rather than the clear discontinuities in the images in 8. We would also expect that when the motion is by different amounts in different parts of the picture, it would be harder to solve the inverse problem.

(4.16) In Davis et al. [1] results are presented for using the rolling shutter technique, but neither the method nor code are provided to explain how. The audio example they provide demonstrates that their rolling shutter technique is not able to record intelligible speech. This technique may well be promising to explore further, if use of high framerate cameras is limited.

4.2 Quality of detected motion and noise

(4.17) Davis et al. [1, equation 8] describe the signal to noise ratio using the formula

$$\mathrm{SNR} = |D_p(\omega)|\frac{\sqrt{n_p}}{\sigma_n},$$

where ω is the frequency, D_p is the amplitude of the motion in the camera image in pixels, n_p is the number of pixels across the image, and σ_n is the standard deviation of the noise. This is as expected, as the "signal" is $D_p(\omega)$ and the "noise" is described by a per-pixel standard deviation σ_n averaged over the number of pixels in the image in the direction of motion (proportional to n_p), giving a standard deviation of the average of $\sigma_n/\sqrt{n_p}$. Typically, they used D_p values between 10^{-3} and 10^{-2}. However, they did not say what signal to noise ratio was necessary to extract meaningful signals. There must also be other important parameters to consider, too, since their results were most sensitive up to 400Hz [1, fig. 7c]. Most results were taken with about a 2kHz framerate and 700×700 pixels. Their attempt with 20kHz and 192×192 pixels seems to have worked worse due to the increase noise (less light per frame) and lower resolution.

(4.18) Since Davis et al. [1] do not investigate different cameras, we turn to the results of D'Emilia et al. [2]. Since they use an inferior algorithm for motion detection, they appear to need several microns of motion for it to be detectable. They used two cameras:
Camera A (AVT Marlin F-131b): 25 fps, 1280x1024px
Camera B (Olympus i-speed): 2000 fps, 1280x1024px

(4.19) The two cameras were mounted in front of a target which vibrates at controlled frequency and amplitudes. The largest displacement was 51mm and 883 m/s^2, in a frequency range of $10 - 2000$ Hz. This paper attempts to find error bounds on the recovered amplitude. In general, they find the uncertainty depends on a large numbers of factors. For camera A, the experimental results showed that the vibration uncertainty is of the order 84 μm, or 3.4% of the amplitude, in the frequency range 10-70 Hz. For camera B, the experimental results showed a vibration uncertainty of 32 μm, 8.4% of the amplitude, in the frequency range 100-300 Hz, while an uncertainty of

13 μm could be achieved (13% of vibration amplitude) in the range 400-600 Hz. In the paper, camera B had an error bound much greater than 10% when considering a low contrast object, for frequencies 300,400 & 500 Hz. Consequently the authors did not analyse this. This could hint at possible problems when analysing video. The paper does not consider object illumination, which could be a large cause of uncertainty for everyday applications.

(4.20) It is unclear how these results should be applied to the MIT technique [1], although it is clear that the detectable signal depends strongly on the characteristics of the camera used. We would therefore propose further experiments using high-speed cameras to investigate the dependence of motion detection on important factors such as the amount of ambient light and the clarity of the image. We propose a thin sheet or strip of some suitable material (such as LDPE) be held vertically, with the top able to be excited horizontally (for example by connecting it to a horizontal loud speaker) and the bottom allowed to move freely, possibly with some added weight at bottom. A high-speed camera would video the motion of the sheet from the side, so that the sheet would appear as a line oscillating left-right in the video. The MIT software would then be used to extract the motion of the sheet from the video. Direct measurements of the oscillation of the sheet could also be made by other means for comparison. Experiments could then be conducted with different cameras, different framerates, different lighting conditions, different camera lenses and distances, and different amplitudes of oscillation, to map out the conditions necessary for successfully detecting motion, and the corresponding noise. This setup could be used with a second loud speaker in the air, allowing the analysis of section 2 to be validated. Moreover, this setup could also be used to investigate the use of rolling shutters (see section 4.1), since each horizontal line of the video will see a slightly different part of the sheet at a slightly different time; this would be particularly interesting when the frame rate of the camera is lower than the natural period of the sheet.

5 Intelligible speech

(5.1) As described in the introduction, voices consist of frequencies from around 80Hz to around 4kHz. Telephone systems use the range 300Hz to 3.4kHz for speech. From their figures, Davis et al. [1] claim to recover intelligible speech using what appears to be only the 250Hz to 850Hz range. In this section, we investigate the requirements for intelligible speech with a focus on this range.

(5.2) This study used list 5 of the "Harvard sentences"[5], which are short sentences each containing 5 important words to be identified. The sentences used were:

[5] `http://www.cs.columbia.edu/~hgs/audio/harvard.html`

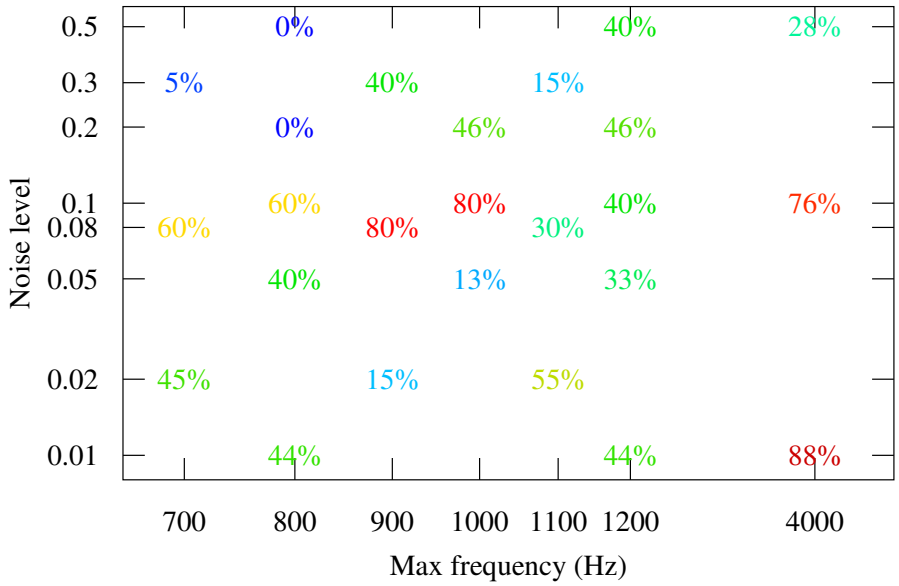

Figure 9: Plot of the average intelligibility (on a percentage scale, with 100% being fully intelligible) of sample sentences with a given lowpass frequency and a given noise level. Based on a very small study of 108 sentences.

1. A king ruled the state in the early days.
2. The ship was torn apart on the sharp reef.
3. Sickness kept him home the third week.
4. The wide road shimmered in the hot sun.
5. The lazy cow lay in the cool grass.
6. Lift the square stone over the fence.
7. The rope will bind the seven books at once.
8. Hop over the fence and plunge in.
9. The friendly gang left the drug store.

Recordings of these sentences were then bandpassed, and white noise added. The upper limit of the bandpass and the amplitude of the noise was varied, with the lower limit of the bandpass set to 300 Hz. In total 12 people each listened to 9 sentences and attempted to identify the 5 important words, leading to an intelligibility score between 0 and 5. The results are summarized in figure 9.

It is clear that a larger sample is needed to get a well-converged average. However, it is also clear that there is a large amount of random variation, with the bottom right corner of figure 9 suggesting that the highest 4kHz cutoff frequency and the lowest noise was necessary to give a repeatably intelligible signal (this being comparable to telephony quality). Based on this, we suggest at a minimum aiming to capture 300Hz to 1.2kHz for a reasonable chance at recognising speech. Not shown in this

figure, but notable from our study, was that native speakers were more able to correctly identify the speech against a noisy background than non-native speakers, even for those non-native speakers with otherwise excellent English.

The results of Davis et al. [1] available to listen to have very little noise, suggesting they have been post-processed to aid intelligibility. Indeed, Davis et al. [1] refer to a number of standard speech enhancement techniques which we have not investigated further here. A description of a possible mathematical basis for enhanced speech recovery using Bayesian inference is presented in appendix A.1.

6 Conclusion

(6.1) This project investigated the feasibility of using the MIT visual microphone technique [1] to recover intelligible speech from high speed video recordings.

(6.2) The MIT experiments [1] depended on loud sound and prior knowledge of "Mary had a little lamb". Their equipment would not have recovered intelligible speech at conversational volume, or if what was being said was not known beforehand.

(6.3) Object oscillations of the order of $0.1\,\mu$m need to be detectable in order to recover speech at conversational volumes. The MIT technique is able to extract motion from videos that at least 1/1000th of a pixel, which gives an indication of the level of magnification of the image needed: at least 10 pixels per millimeter, and preferably 100 pixels per millimeter.

(6.4) The investigation in section 5 suggests that, for intelligible speech, we would need to capture at least up to 1.2 kHz, and preferrably higher (potentially 3.4 kHz for telephony quality sound). The MIT technique is able to detect frequencies up to about 850 Hz using a 2,200 fps camera, suggesting that at least a 3,100 fps camera is needed, and potentially a 9,000 fps camera. However, when the MIT researchers tried higher framerates, they found the increased noise made intelligibility harder. Clearly understatnding the signal to noise ratio of the captured sound is important.

(6.5) While we would have liked to be able to give explicit advice about the hardware requirements for obtaining intelligible speech using the MIT technique, to do so would need further experiments. The model of signal to noise ratio in the MIT paper [1] is experimentally derived, and does not account for changes in ambient lighting or camera framerate, camera pixel noise (ISO setting), image clarity (e.g. caused by imperfect camera lenses), nor does it describe the frequency dependence of the signal to noise ratio. A suggestion for a possible future experiment is described at the end of section 4.2.

(6.6) Whatever the specific requirements, a high speed, low noise camera is essential. Existing video footage is unlikely to be of sufficient quality or sufficiently high frame rate.

(6.7) While MIT report using the rolling shutter of a camera to capture frequencies much higher than the camera frame rate, this did not result in intelligible speech (or even in identifiable speech), and the MIT algorithm for doing this has not been made public. While a much more complicated and challenging problem, we see no fundamental reason why rolling shutters could not be taken advantage of if access to high speed cameras is limited.

(6.8) The MIT experiments depend on the motion of light objects (such as crisp packets or disposable cups), but resonances of heavier objects (e.g. curtains) could potentially also be exploited (section 2.4. Due to the drop in oscillating amplitude with increasing frequency (figure 2), it is likely that only the first few resonances will be useful, which could simplify the task of taking advantage of resonances; the MIT results are also suggestive that only the first few modes are important [1, figure 5]. The closer together the resonances are, the larger the resulting motion (figure 3).

(6.9) By looking at reflections in objects, much more subtle motion could be detectable, possibly including speech at conversational volumes. For example, using reflections in a bending 10 cm mirror viewed from 10 m gives a 400× magnification of motion.

(6.10) Vibrations of the camera were ignored by MIT, but are expected to be significant. The MIT technique could probably be adapted to be resilient to camera vibrations, by using the lowest resolution wavelets as a proxy for whole-scene motion, although this was not attempted here.

(6.11) We are sceptical of the MIT claim that optimized computer code could process video in real time. The technique is highly parallelizable, however, so could possibly be offloaded to a supercomputer for real-time processing.

(6.12) The effect of video compression artifacts was not considered in this work, nor by MIT. Modern video compression (such as H.264 and H.265 codecs often used in mp4 files) use motion estimation to enhance compression, and the motion estimation used is unlikely to be sufficiently accurate to recover the subtle motions needed for the MIT technique.

(6.13) Also not considered here or by MIT is aliasing, where frequencies higher than half the camera frame rate *alias* and become artifacts at lower frequencies. In audio recording, it is essential to use a low-pass filter before digitising the sound in order to eliminate aliasing, but this cannot be done with current high-speed cameras.

(6.14) It may be possible to build cheap counter-measures to counter this technique. For example, small oscillators (such as the vibration units in mobile phones) could be stuck onto surfaces and generate small amounts of white noise oscillations. These would be inaudible to people in the room, but would render the visual microphone technique impossible.

A Appendices

A.1 Recognising speech from a noisy background

(A.1.1) Inferring speech from the minuscule vibrations in video would be enhanced if one had a strong prior probability distribution on speech. Singing is well described as filtered white noise. Perhaps speech could be too. The filter is a function of time, to make the changes in phonemes, pitch, loudness, speaker etc. One could model this by a Markov process.

(A.1.2) Suppose the sound source x is given by

$$\dot{x} = Gx + Lu \tag{35}$$

where G is an asymptotically stable linear map, u is a vector of unit white noises, and L is a linear map.

(A.1.3) Model the response y of the objects in the sound field as a linear system

$$\dot{y} = Fy + Mx + Pv \tag{36}$$

where F is another asymptotically stable linear map, M a linear map, v some more white noises, and P a linear map.

(A.1.4) Model the observations z by

$$\dot{z} = Kz + Hy + Nw \tag{37}$$

with K asymptotically stable, H, N linear, and w more white noises. One could take z to be an instantaneous measurement, e.g. $z = -K^{-1}Hx + \varepsilon$ as would result from K being large, but in general one might expect some correlations between the measurement errors ε and the above formulation gives some.

(A.1.5) Then the full system is

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} G & 0 & 0 \\ M & F & 0 \\ 0 & H & K \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} L & 0 & 0 \\ 0 & P & 0 \\ 0 & 0 & N \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}. \tag{38}$$

The problem is to infer the function x from the function z. This is a Kalman filter problem and has a standard efficient solution. The matrices are unknown, however, so they would also need to be inferred. There is some redundancy in the description, e.g. for the matrix, call it C, in front of the white noises, only the matrix CC^T matters, and one could apply linear coordinate changes to x and y, so before attempting to infer the matrices it might be best to normalise them. Furthermore, it is likely

that only the lightest damped modes of F matter, so one could attempt a highly reduced description of F.

(A.1.6) Extension to allow the filter G and noise amplitudes L to vary slowly in time takes one out of the autonomous regime for the efficient Kalman filter solution, but it is still a Gaussian process so may be relatively feasible to do the inference.

References

[1] Davis, A., M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand, and W. T. Freeman (2014). The visual microphone: Passive recovery of sound from video. *ACM Transactions on Graphics (Proc. SIGGRAPH) 33*(4), 79:1–79:10.

[2] D'Emilia, G., L. Razzè, and E. Zappa (2013). Uncertainty analysis of high frequency image-based vibration measurements. *Measurement 46*(8), 2630–2637.

[3] Landau, L. D. and E. Lifshitz (1970). *Theory of Elasticity* (2nd ed.). Pergamon.

[4] Portilla, J. and E. P. Simoncelli (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comp. Vis. 40*(1), 49–71.

[5] Simoncelli, E. P. and W. T. Freeman (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. 2nd IEEE International Conference on Image Processing, Washington DC*, Volume 3, pp. 444–447.

[6] Simoncelli, E. P., W. T. Freeman, E. H. Adelson, and D. J. Heeger (1992). Shiftable multi-scale transforms. *IEEE Transactions on Information Theory 38*(2), 587–607.

[7] Wadhwa, N., M. Rubinstein, F. Durand, and W. T. Freeman (2013). Phase-based video motion processing. *ACM Trans. Graph. (Proceedings SIGGRAPH 2013) 32*(4).