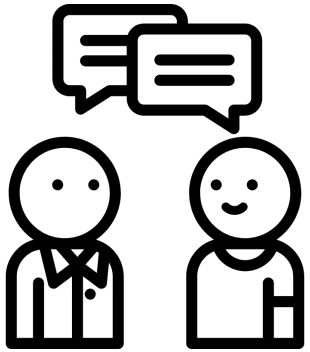
The Visual Microphone: Passive Recovery of Sound from Video

Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo Durand, William T. Freeman

GDL 22 avril 2026
Gaspard Thévenon

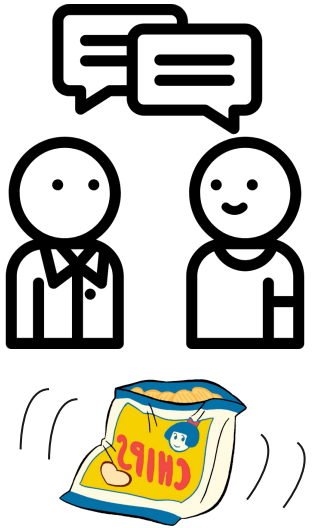
How can I spy on my neighbors ?

Situation: you want to retrieve the sound somewhere too far to hear.



How can I spy on my neighbors ?

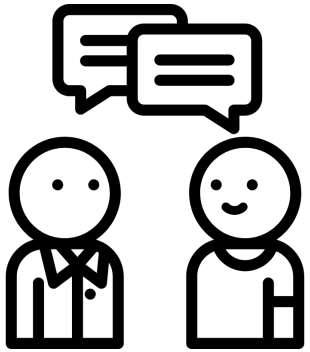
Situation: you want to retrieve the sound somewhere too far to hear.
Solution: Measure the vibrations of surrounding objects.



How can I spy on my neighbors ?

Situation: you want to retrieve the sound somewhere too far to hear.
Solution: Measure the vibrations of surrounding objects.

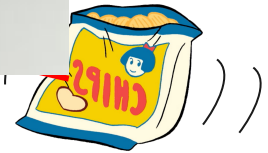
Laser Doppler vibrometer



How can I spy on my neighbors ?

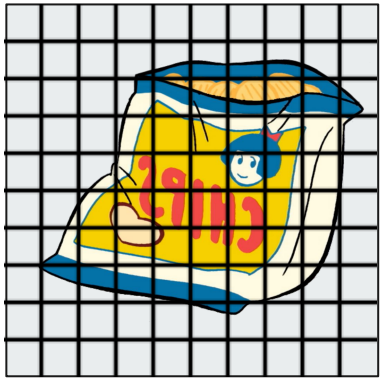
Situation: you wan

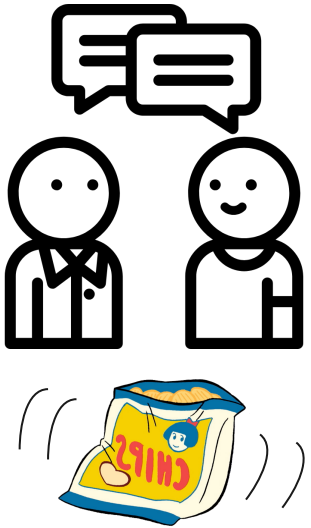
Solution: Measure



How can I spy on my neighbors ?

Situation: you want to retrieve the sound somewhere too far to hear.
Solution: Measure the vibrations of surrounding objects.



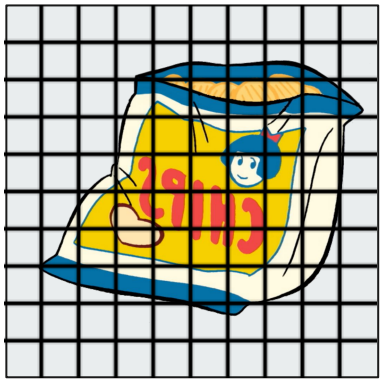


How can I spy on my neighbors ?

Situation: you want to retrieve the sound somewhere too far to hear.
Solution: Measure the vibrations of surrounding objects.

?

$$$, 5k FPS

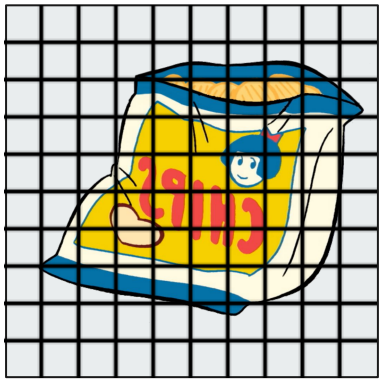


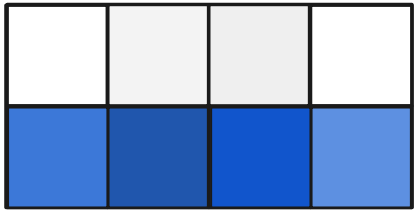
Two limiting quantities:

- Frequency of vibrations / Camera FPS
- Amplitude of vibrations / Camera resolution

  Typically, ~ 1/100th of a pixel

Measure sub-pixel vibrations





Idea:

vibrations are too minute to “skip” pixels

but they lead to small intensity variations at the object borders

however, this measure is very noisy (e.g. camera sensitivity) and is unusable in practice

Measure sub-pixel vibrations

Video: $I(x, y, t)$

If we take a frame: $I(x, y)$

It can be seen as a 2D spatial signal.

For a fixed frequency ω_s and orientation θ, we apply a Gabor filter.

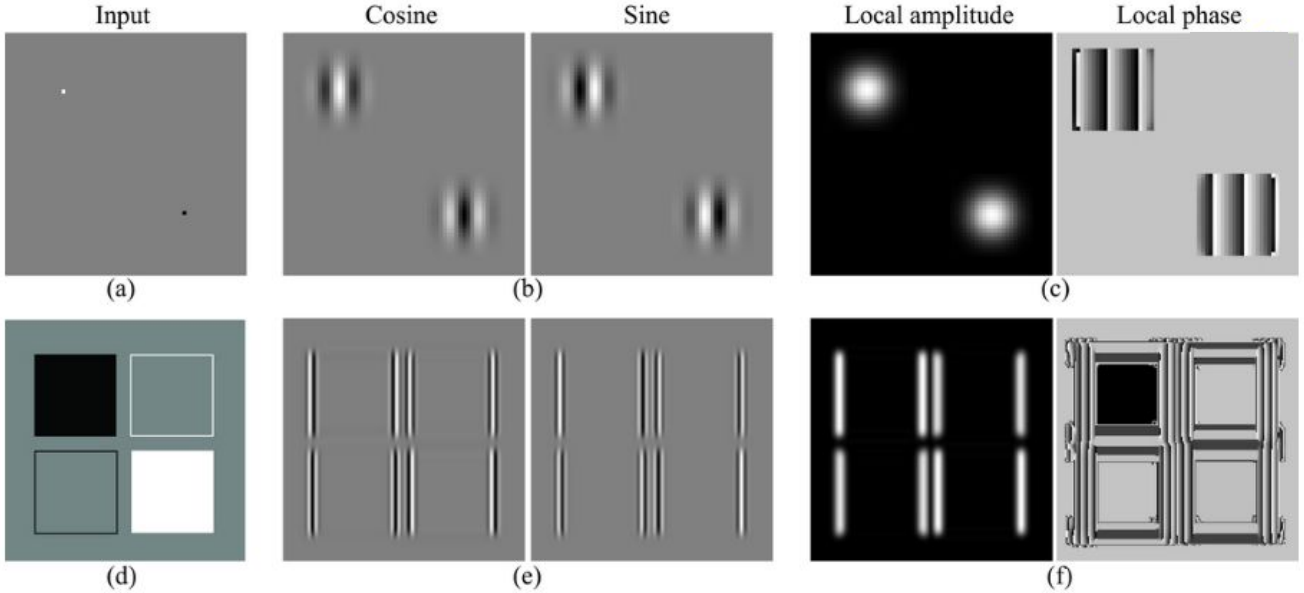
$$g_{s,\theta}(x, y) = \frac{1}{2\pi\sigma_s^2} \exp\left(-\frac{x_\theta^2 + \gamma^2 y_\theta^2}{2\sigma_s^2}\right) \cdot \exp(i\omega_s x_\theta)$$

where (x_θ, y_θ) is the rotated image plane.
Convoluting the image with this filter gives a complex-valued image:

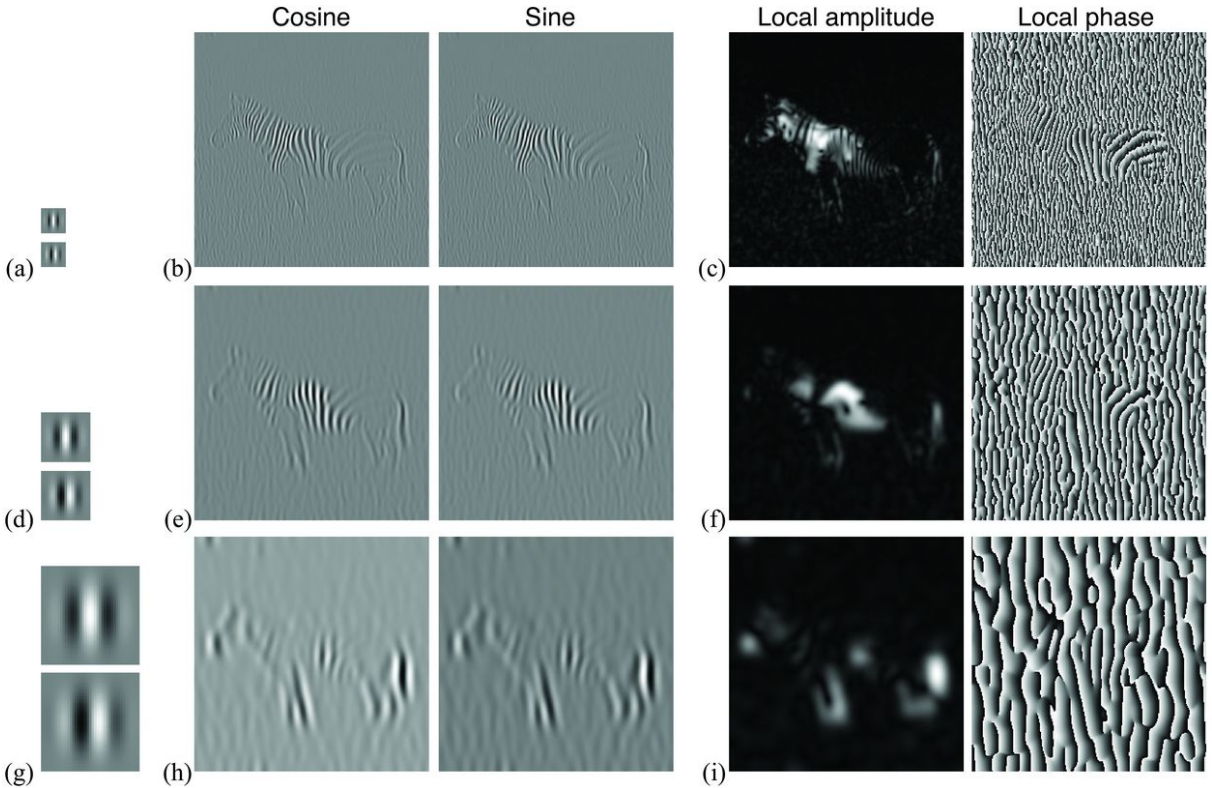
$$(I * g_{s,\theta})(x, y) = A_{s,\theta}(x, y) \exp(i\phi_{s,\theta}(x, y))$$

Amplitude: “How much does this periodic pattern appear here?”
Phase: “Where are we located in this pattern?”

Measure sub-pixel vibrations



Measure sub-pixel vibrations



Relationship between phase and vibration

By doing the analysis for every frame, we obtain the temporal variation of the phase: $\phi(x, y, t)$

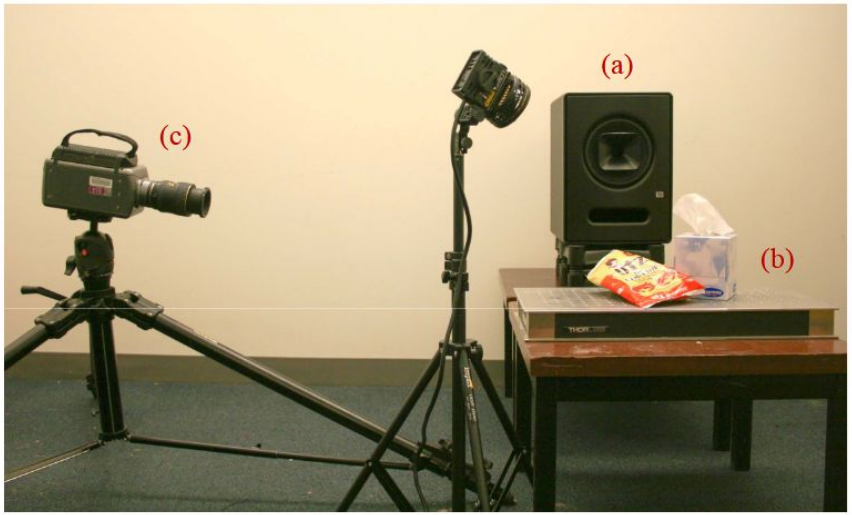
Locally, for a small temporal displacement, the amplitude varies slowly, but the phase varies quickly:

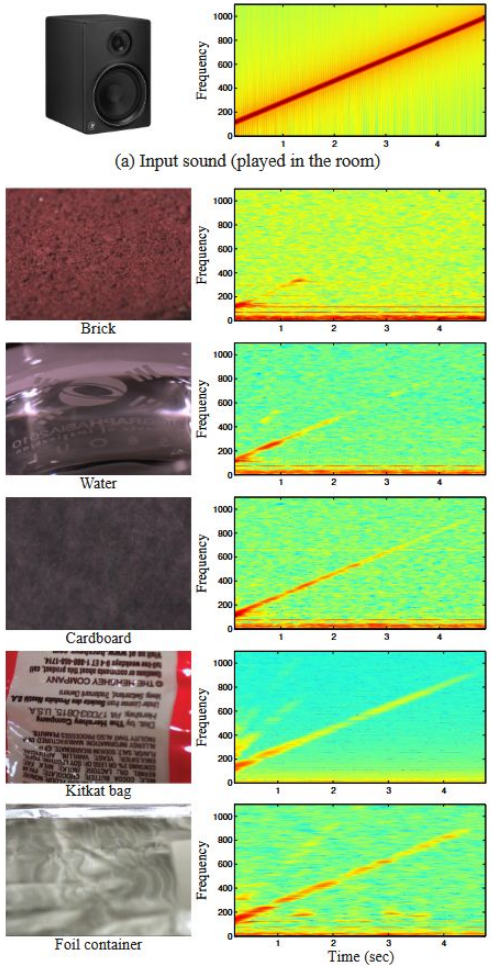
$$\Delta\phi(x,t) = \phi(x,t) - \phi(x,0) \approx -\omega_s \delta(t)$$

So the phase variation is linearly correlated to the object vibration.

The approach is based on this idea: recover the local phase variation (using a pyramidal filter bank), which is a much more reliable, and then recover the object vibration, and thus the audio signal.

Which material is better ?





Which material is better ?

crabchips

greenteabox


tissue


foiltogo

kitkat


afoil

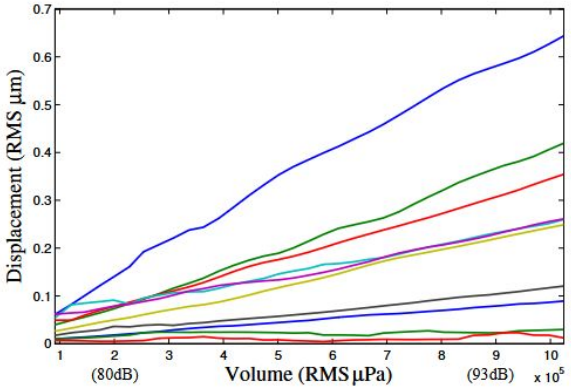

rose

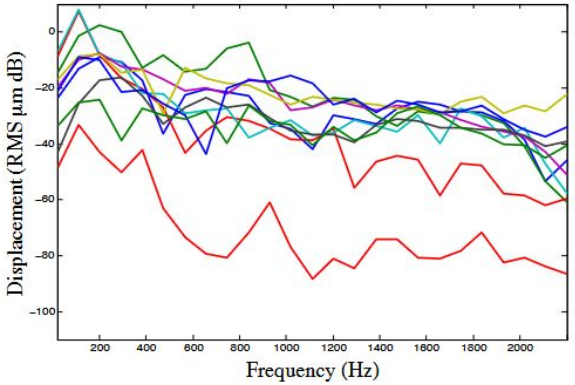

foamcup

chobani

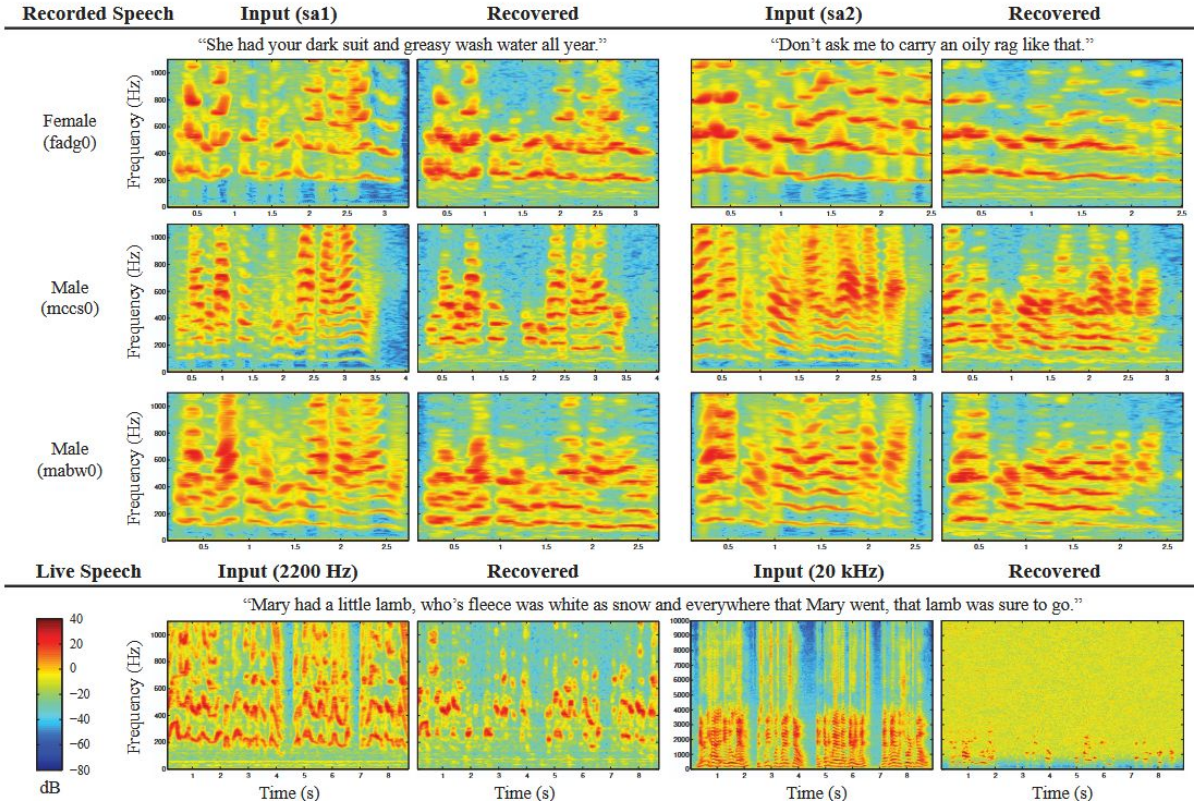

teapot

	Measured displacement increase rate (μm/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

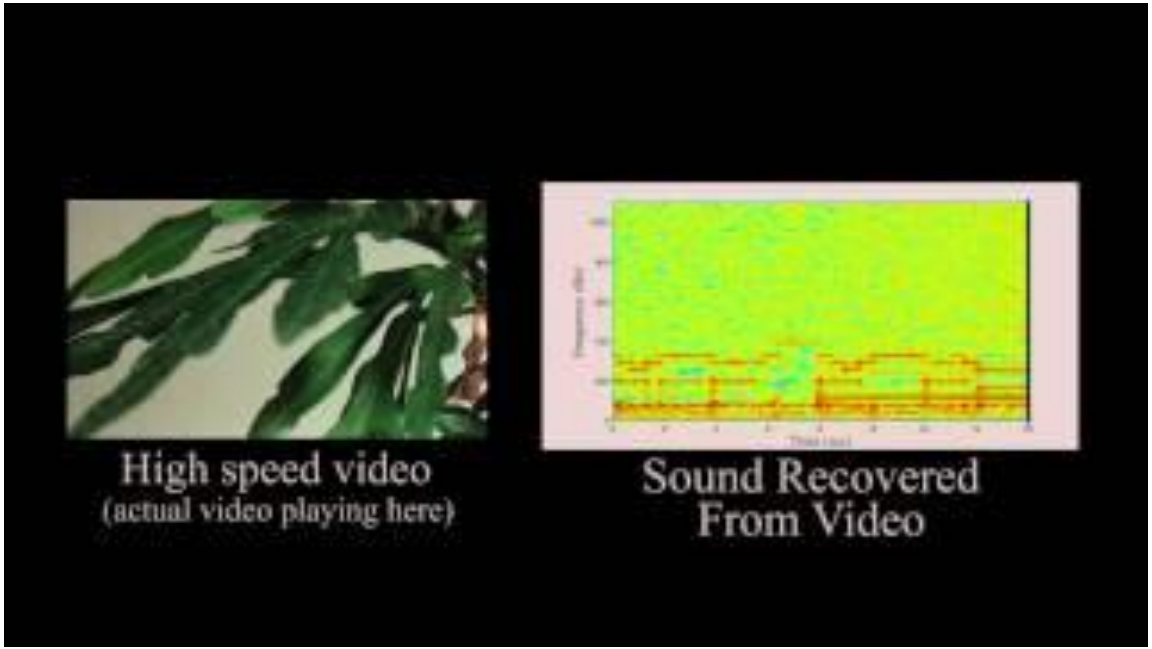




Speech reconstruction #CIA



Examples


High speed video
(actual video playing here)
Sound Recovered
From Video

Thank you, questions ?