Technical Report Brief

of a Research Article

The visual microphone: passive recovery of sound from video [1]

Faculty of Engineering, Ariel University

Department of Electrical Engineering

Written by: Ronel Herzass

Lecturer: Mr. Nadav Goldman

Date: 20/10/2025

Table of Contents

Abstract

Sound waves travel through mediums and cause vibrations on the surface of objects. In this article the writers described using a high-speed video camera to extract the vibration from different, everyday surfaces to recover the sounds that were produced from those vibrations. The writers used both real and simulated data to examine the properties that enable them to recover sound using vision. The research compared the quality of the recovered sound using intelligibility and SNR (sound-noise ratio) metrics. The rolling shutter effect in consumer cameras was also used to recover sound from vibration on surfaces. The results show that it is possible to recover speech and music from inside a room using a video analysis of a bag of chips. The research shows a method to convert everyday objects into potential microphones.

Keywords

Remote sound acquisition, sound from video, visual acoustics

Introduction

Sound waves are the oscillation of pressure in a medium. When sound waves hit an object, they cause the surface of that object to vibrate. These vibrations create visual signals by deforming the surface of the object.

The objective of the study was to prove there is a passive way to recover sound using video. The motion signal was extracted and characterized. In addition, a demonstration of the rolling shutter effect in a CMOS sensor was shown as an example of a method to recover sound from standard frame-rates cameras [2][3].

Methodology

General Process

Video input was decomposed into spatial sub-bands corresponding to distance and orientation values using a complex steerable pyramid filter [4]. Every local phase is evaluated compared to the first frame to compute the phase vibration. Local motion at every pixel was computed and combined to produce a single global motion of the object.

Denoising was done by improving the SNR (Sound to Noise Ratio) with a high-pass Butterworth filter applied on the signal. For intelligibility applications the signal was processed through a speech enhancement algorithm [5] that takes into consideration human perception of speech.

The experiment was done indoors using a high-speed camera, loudspeaker, and an examined object. The speaker was placed on a stand separated from the object to avoid interference. The experiment was done with a photography lamp and at distances ranging from 0.5 to 2 meters, and in daylight from outside a glass door.

First, a test for recoverable frequencies was done by playing a linear ramp of a sine wave with linearly increasing frequencies at a variety of different objects. A brick was used as a control experiment because of its rigidness and weight. Then, another experiment reconstructed sound for several human speech samples from the TIMIT dataset as well as live human speech.

Another method to recover sound was by using the rolling shutter in the CMOS sensor of the video camera, where the rows in the sensor are captured at different times. Sound was recovered from each row instead of from each frame by knowing key parameters of the sensor and calculating the motion signal reflecting the audio signal.

Required equipment

1. Phantom V10 High Speed Camera
2. Loudspeaker
3. Ordinary recovering testing object

Results

The experiment showed that sound can be recovered from many different everyday objects at high intelligibility, as shown in the following table. A comparison recovering speech samples from a bag of chips using the method in the experiment (VM) and a Laser Doppler vibrometer (LDV).

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 - Comparison of VM and LDV methods for sound recovery

It was also shown that in high frequencies, the recovered signal is weaker, as expected, because high frequencies produce smaller displacement and are reduced by most materials. Furthermore, higher frequencies in lighter objects were easier to recover than in inert objects.

Discussion

The study attempts to recover sound input by monitoring motion in nearby objects. Analyzing the response of an object from a nearby sound shows that there is a linear connection to the motion of the object, which models the system as a linear time invariant system and can be easily monitored. Recovering unintelligible sounds can be also useful to identify the number or gender of speakers in a room in surveillance scenarios.

The method examined in the study is limited by the magnification of the lens. Key parameters that are used to recover sound are proportionate to the number of pixels which increase with the magnification lens. Recovering sound for larger distances may require expensive optics with large focal lengths.

The ability to reconstruct sound from the vibrations of innate objects opens numerous real-life uses such as reconstruction of speech for intelligence purposes or personal safety by using street cameras to monitor car windows and avert infant car deaths.

Conclusions

The study shows that visual vibration in objects caused by sound can be backtracked and reconstructed to recover the original sound. Sound recovery from everyday objects may lead to uses in a variety of fields, such as aiding the hearing impaired, surveillance and espionage, and understanding patterns in nature.

Further research will have to address both ethical and legal issues, such as invasion of privacy and the use of unauthorized information. Furthermore, awareness must be heightened once this possibility becomes more practical and extensively available.

References

[1] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo 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. https://doi.org/10.1145/2601097.2601119

[2] Junichi Nakamura. 2005. Image Sensors and Signal Processing for Digital Still Cameras. CRC Press, Inc., USA. https://dl.acm.org/doi/10.5555/1211284

[3] 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.

[4] Portilla, J., Simoncelli, E.P. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision* 40, 49–70 (2000). https://doi.org/10.1023/A:1026553619983

[5] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," in *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857-869, Sept. 2005.