Digital Semaphores: Voice Recognition using Visual Microphones

Prahit Yaugand

Abstract

Slow motion videos render a new perspective to mundane actions, like pouring a glass of water or throwing a water balloon. They convey emotion into a scene and are enticing and impactful. A typical slow motion video records at 60 frames per second. With current advancements in technology, high-speed cameras have been able to provide insights into the visualization of sound. This paper explores a novel approach to recover audio from high-speed video recordings. Visual recordings of vibrations are analyzed to regenerate the original audio wave. Word recognition and classification is performed successfully by a convolutional neural network that is trained using spectrograms of the recovered audio files.

1. Introduction

Recent advancements in optical technology led to cameras with multiple features, such as the ability to capture faster movements and zoom in to view microscopic details. Over the last few decades, powerful algorithms have emerged in computer vision software that facilitate locating patterns across frames. Convolutional neural networks have employed mathematical operations to classify and identify images. Put together, these technologies can be used to capture vibrations visually and recover audio from a recording.

Sound waves propagate by oscillation of particles on their path. These vibrations are so minute that it is difficult to notice them with the naked eye. Only sensors tailored to this task can acquire these microscopic vibrations. Sound sensors are typically composed of a thin membrane placed in the sound wave's path. Sound causes this membrane to vibrate, and the sensor's transducers convert those fluctuations into electrical signals of corresponding intensity. This paper analyzes an innovative technique to visually record vibrations of a surface by capturing a series of images. The images are then analyzed to recreate the sound wave and identify the spoken words using a convolutional neural network. After training, the neural network provided an accuracy of 97.33% on validation data.

The techniques used in this study have several applications, ranging from the detection of biological rates such as heart rate and respiratory rate to espionage. The recreation of sound from a visual vibration is accomplished through the series of steps as described below:

1. Silent Video Generation: Vibration videos of amplified sound are recorded and captured for the digits from 0 - 9.
2. Movement Tracking: Regions are chosen and tracked through each frame to ascertain the movements of the surface
3. Audio Recovery: A region with high standard deviation, present in most of the frames, is chosen and its movement through frames is computed and converted to an audio file, which is then plotted as a spectrogram (frequency-time-intensity plot)
4. Spectrogram Classification: The spectrograms are fed to the neural network where they are classified into numbers from 0 - 9

The rest of the paper is organized as follows: Section 2 discusses several instances of related work and provides a brief overview of multiple close research ideas. Section 3 depicts the method, experimental setup, the data and the software. Section 4 discusses the results. Section 5 presents the conclusion and future work.

2. Related Work

Human brains are capable of sensing sound much faster than sight. The human visual system has a low temporal resolution and high spatial resolution whereas the auditory system has a high temporal resolution and low spatial resolution [17]. Technology has evolved to serve this differentiation. Audio equipment has a higher standard sampling rate (44100 fps) whereas video equipment captures frames at the rate of 30 frames per second. To recover audio from video, the rate of sampling for video needs to be increased by a factor of magnitude. High-speed cameras [more details in Appendix A.3], equipped with extremely fast shutter speeds, can observe swift movements and record them in high temporal resolution. High-speed cameras have continued to improve and become affordable in recent years, coinciding with the rise of computing techniques for visual data processing and pattern matching leading to a rise in the number of studies in this area.

The Visual Microphone: Passive Recovery of Sound from Video [1] Researchers recovered sound from high-speed video footage of a variety of objects with different properties by analyzing the change in vibrations throughout each object's surface. Analyzing the variation of the sound through different locations allowed these researchers to recover a sound similar to the original audio.

Lamphone: Real-Time Passive Sound Recovery from Light Bulb Vibration [4] Audio was obtained from a high-speed video of a lightbulb's vibrations. This work uses a remote electro-optical sensor to analyze a lightbulb's fluctuations due to audio. By using these fluctuations, researchers recovered speech which was successfully identified by the Google Cloud Speech API.

Dual-Shutter Optical Vibration Sensing [5] A laser traced a speckle pattern on an oscillating surface. This pattern amplified the object's vibration and allowed for higher-quality sound recovery.

3. Method

This section consists of the experimental procedures and delves into aspects from the dataset creation to the audio classification.

3.1. Silent Video Generation

A loudspeaker played audio in front of a surface, resulting in minute vibrations, which were captured by a high-speed camera. These recorded videos were labeled based on the audio file for classification.

3.1.1 Setup

A high-speed global shutter camera, the Freefly Wave, was chosen with the Venus Optics Laowa 24 mm lens for this study. As the surface vibrations were critical for this research, thin materials like silk, polyester, paper, and cotton along with variants of multiple substances were evaluated. The surface was hung in front of a speaker to record minute oscillations. When the audio was played in the speaker, the air in front of the speaker vibrated, inducing the surface to oscillate. Videos were recorded on multiple surfaces, to evaluate their sensitivity to sound and the quality of the oscillations that could be observed. Various sizes, colors, and patterns were experimented with and it was discovered that effective recovery of sound occurred in white paper with black dots. The speckled paper provided observable movement as the black dots highly contrasted with the white background allowing the template matching function to classify the exact position of the dot at each frame. A picture of the complete setup is shown in Figure 1.

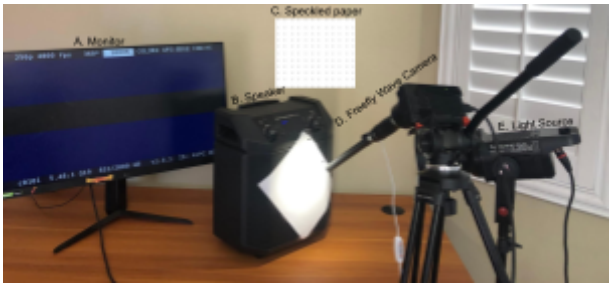

Figure 1: Video recording setup and vibrating surface

High-speed video cameras had several limitations. By definition, capturing at high speed implied that the optical sensors in the camera were exposed for a fraction of a second. Intense lighting was needed to capture each frame. Shooting at 4800 frames per second requires around 150 times more lighting than 30 fps. When the shutter opens, light travels through the aperture of the camera to reach the lens. The lens focuses the light on a digital sensor, which saves the image to a hard drive. In a low-lighting environment, the image appears dark and cannot be observed well. This study used the Aputure Light Storm 120D-Ⅱ, which allowed for the capture of videos at 4800 frames per second.

3.1.2 Video Dataset Capture

A set of pre-recorded human voices (digits 0 - 9) were played through a speaker. A surface (a paper printed with multiple tiny dots) was placed in front of the speaker. Sound waves travel in the form of compressions and rarefactions in air which causes damped vibrations on the surface: the dots would move forward and backward. An observer with a camera placed directly ahead would not be able to perceive the vibration due to lack of depth perception. Humans can perceive depth by observing the same scene through two eyes. To overcome the lack of depth perception and to view the motion, the camera should be placed at an angle as shown in Figure 2. A detailed proof relating to the placement of the camera is discussed in Appendix A.2 for reference.

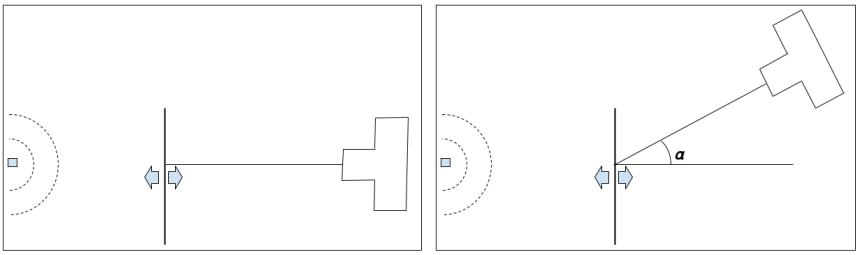

Figure 2: Camera viewing object

A. Camera facing surface horizontally shown on left B. Camera at an angle to surface shown on right

The high-speed camera recorded the vibrations at 4800 frames per second. When the video was played back at 30 fps (160x slower), the visual vibrations of the dots on the surface could be perceived with the naked eye. Recording these rapid fluctuations was essential for the recreation of the sound.

In summary, a recorded sound was amplified through a speaker to create vibrations on a surface placed in front of the speaker. Enough lighting was provided for a high-speed camera to track any minuscule motions. The recorded videos were marked with labels on their filename to be used by the convolutional neural network model for classification. A pictorial representation of the silent video generation is shown below in Figure 3.

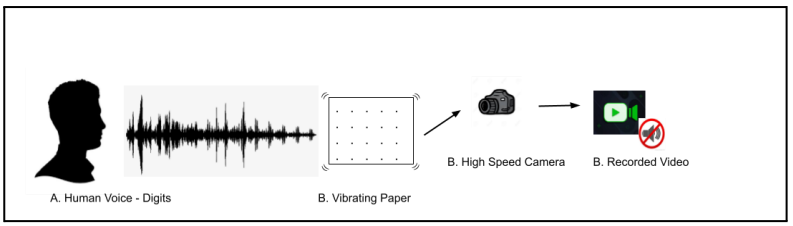

Figure 3: Silent Video Dataset Creation Process

3.2 Movement Tracking

Standard preprocessing techniques were used across all images in this study. Images were converted from RGB to grayscale to reduce computational requirements. Then the images were normalized to achieve higher contrast between neighboring pixels. A high contrast area in the first frame of the video was chosen with a configured size (around 50 x 50) and identified as a template. Once a template was chosen, matched regions located in the further frames help track the minute motions of the vibrating surface. A detailed process is shown in Figure 4.

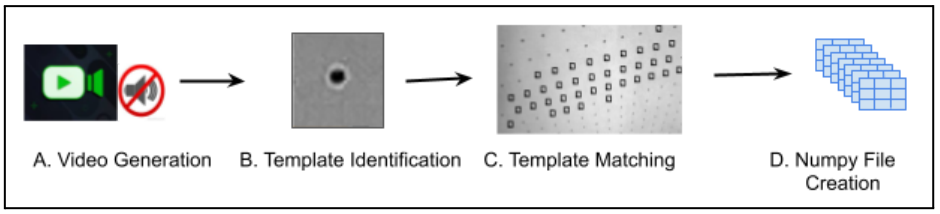

Figure 4: Movement Tracking Process

3.2.1 Choosing Key Areas

Frames in the video were cropped to focus on the moving surface to ensure that the background did not interfere with the result. The image was split into multiple smaller sections as configured, and the standard deviation for each section was computed. The section with the highest standard deviation, representing the portion of the image, with the maximum variance among its pixel values, was referred to as the template and was chosen to track vibrations. As the surface chosen for the study was a paper with dots, templates consisted of a black dot on a white background. Due to the disparity in colors, even slight movements could be detected and traced by the camera. When viewing a monochromatic surface, it is difficult to distinguish movement, but in this scenario, the dot's movement allowed the motion (shown in Figure 5) to be recorded with greater accuracy.

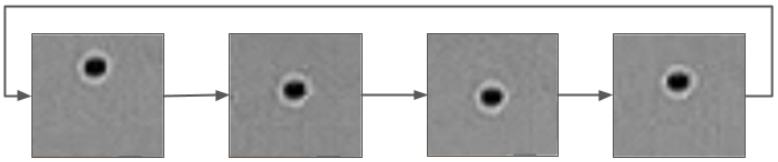

Figure 5: Plotting the movement
Periodic motion of a dot as it vibrates through the video

3.2.2 Tracking Movement Across Frames

After template selection on the first frame, further frames were analyzed using the original template. The normalized correlation coefficient was calculated for the input image in the frames ahead. Positive values were accepted as matches and negative ones were rejected. The formula for the

coefficient is displayed in Equation 1. Accepted regions in all frames after the first were sent for further processing.

$$R_{ccoef\_normed} = \frac{\Sigma_{x',y'} T'(x',y') * I'(x+x',y+y')}{\sqrt{\Sigma_{x',y'} T'(x',y')^2 * \Sigma_{x',y'} I'(x+x',y+y')^2}}$$

$$T'(x',y') = T(x',y') - \frac{\Sigma_{x'',y''} T(x'',y'')}{w-h}$$

$$I'(x+x',y+y') = I(x+x',y+y') - \frac{\Sigma_{x'',y''} I(x'',y'')}{w-h}$$

Equation 1: Normalized Correlation Coefficient Matching Method

R: normalized correlation coefficient, T is the template and I is the input image

Matched templates were grouped to avoid duplicate template identification. To do this, grouped templates' center points were averaged and the mean x and y coordinates were saved as a center. This method ensured that the recorded coordinate was selected through the same means each time. Overall, as each template was visualized through several frames, the movement of each high-variance location was analyzed accurately.

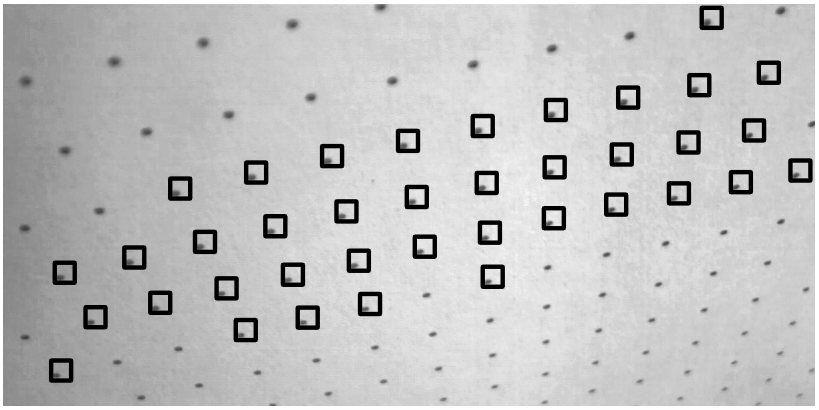

Figure 6: Template Matching Process

Each square depicts a matched region in a frame

Once all of the frames were scanned, the frame number and the change in x-coordinate and y-coordinate for each match (example shown in Figure 6 above) were recorded in a numpy file.

3.3 Audio Recovery

The numpy file was used in the creation of a graph with the y-coordinate against the frame. Filtering and normalization were performed on the graph before reconstructing the audio to reduce noise. The audio file was then converted to a spectrogram and fed into the model. This method is shown diagrammatically in Figure 7.

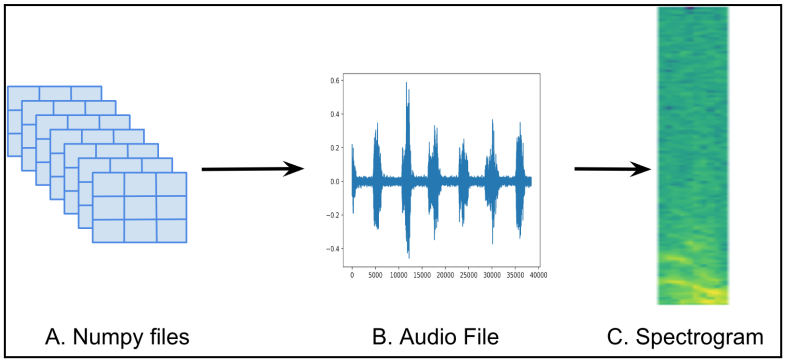

Figure 7: Spectrogram generation

3.3.1 Recreation of Audio

To plot the data, the frame number was divided by the video's fps to get the time axis for each frame. For example, 500 frames taken in a video of 1000 frames per second would last half a second. Plotting the time against the y coordinate of the image allowed for the visualization of the movement of the specified region. The audio was then normalized to record relative changes in the periodic motion. A low pass filter was placed to remove frequencies above the Nyquist frequency (discussed in Appendix A.1) as aliasing frequencies could pose as hidden low frequencies. A high pass filter was subsequently applied that removed frequencies below 20, eliminating noise caused by the vibration of the paper. After applying these filters, the audio file was recreated at a sampling rate of the video's fps. The audio could then be played through any audio player.

3.3.2 Spectrogram

A spectrogram is a visual representation of an audio file. A spectrogram has three axes: (1) time on the x-axis, (2) frequency on the y-axis, and (3) intensity on the z-axis - shown as a color range in Figure 7 above. Spectrograms are obtained in many situations, for example, to remove unwanted noise in a signal, identify spoken words phonetically, and analyze real-world data with various frequency components. These plots are used in the fields of linguistics, music, sonar, and speech processing. This study utilizes spectrograms to classify recovered audio as a specific word. Since each spoken word has different intonations varying over time, spectrograms are useful in this scenario. The spectrogram is resized into a 256 * 32 array to keep image sizes identical and is fed into the model. When developing the spectrograms, the audio files were first split into individual number parts. Then each part was converted into a spectrogram.

3.4 Spectrogram Classification

This study allowed for substantial recovery of the original audio so that words could be recognized from the recovered audio. However, there was noise present in the recovered signal. In order to accurately identify recovered audio files, a convolutional neural network was designed to classify recovered audio files into numbers as shown in Figure 8.

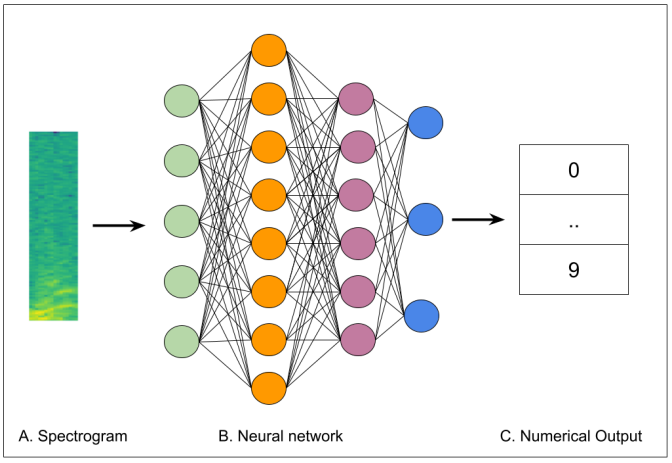

Figure 8: Audio Recognition

3.4.1 Model

Among the various deep learning neural networks, convolutional neural networks, also known as convnets are popular for their effectiveness in identifying patterns across images. A convnet is a stack of a convolutional layer and a pooling layer. The first segment of the neural network employed a sequential model with three such stacks. The second segment utilizes a densely connected classifier network with a stack of dense layers. A flatten layer connects the three dimensional output from the convnets to the one dimensional input in the classifier network. The model applied batch normalization after every layer to regularize and normalize the data. A rectified linear unit (ReLU) was used as the activation function in every layer due to its fast and straightforward computational strength. The final dense layer reduces the output to 10 classes, matching the training classes from 0 to 9. A diagram of the model is shown below in Figure 9.

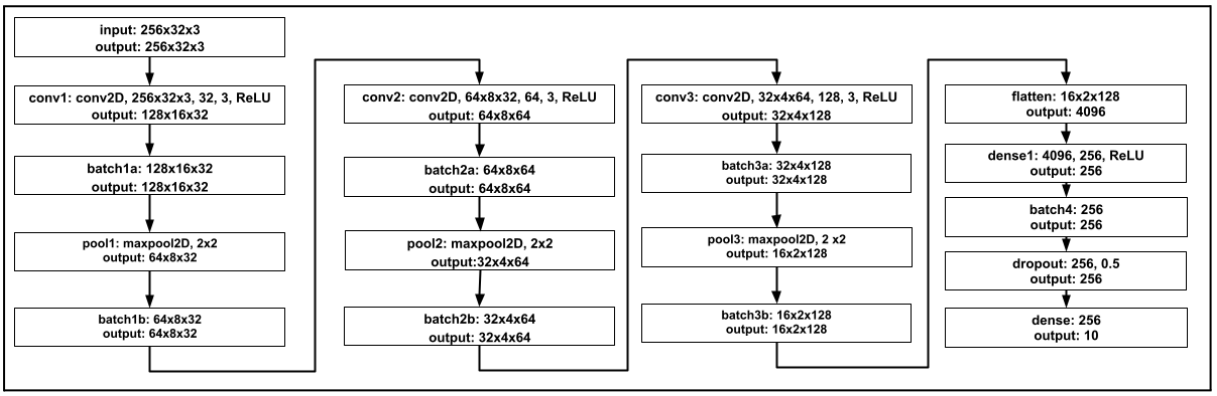

Figure 9: Model

Diagrammatic display of model

4 Results

This section describes the results of the study in multiple steps. The configuration of the camera, lens for slow motion video recording, and high lumens lighting were critical parts of the video generation process. After the camera configuration was set to take a 2048 * 256 * 3 pixel video at 4800 frames per second with necessary lighting, multiple silent videos were recorded for the digits 0 - 9.

The video processing unit converted the videos to grayscale and normalized them to make the image sharper and highlight the contrast between the black dot and the white background. During the

template creation and the template matching process, in order to acquire precise plots, templates that were absent from most of the frames were eliminated, and the remaining were sent to the subsequent processes as a numpy file. To eliminate noise, filtering was performed on the data and the audio file was recreated as a wav file. It was difficult to identify the audio, but individuals familiar with the different original audio files could discern similarities between the original and recovered audio. The audio was then converted into a spectrogram which allowed the convolutional neural network model to classify the audio files with a validation accuracy of 97.33%

4.1 Dataset

Three different datasets were created during this study.

(1) High speed video dataset
(2) Recovered audio wave dataset
(3) Recovered audio spectrogram dataset

The high speed video dataset contains 1000 labeled samples that map a video file to a spoken number. All of these videos were recorded at the speed of 4800 frames per second. The recording spans a timespan of around 0.3 seconds which is the time taken to speak a number. The second dataset contains the mapping for the recovered audio waves to the spoken number. These files are much smaller than their corresponding video files. The third dataset consists of a total of 1000 labeled samples spread across the numericals of 0 through 9. Each sample contained the spectrogram of the reconstructed wave of visual vibrations of the speckled paper when a number from zero through nine was spoken and its corresponding label was the number that was spoken.

4.2 Audio Plots

The normalized audio file plots of frames vs relative intensity is shown in Figure 10 for few of the spoken numbers. Each audio plot has the frame number on the x-axis and relative intensity on the y-axis.

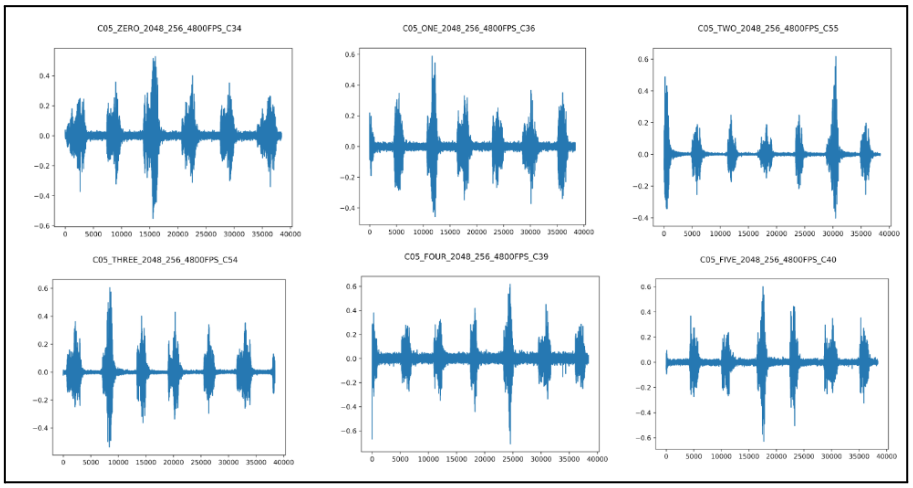

Figure 10: Audio Visualization of multiple digits 0s - 5s in order (left to right, top to bottom)

4.3 Spectrogram Plots

The spectrograms for the audio files are presented in Figure 11 for the digits 0 - 9. Minor differences can be seen between different numbers. This is what allows for successful classification by the model.

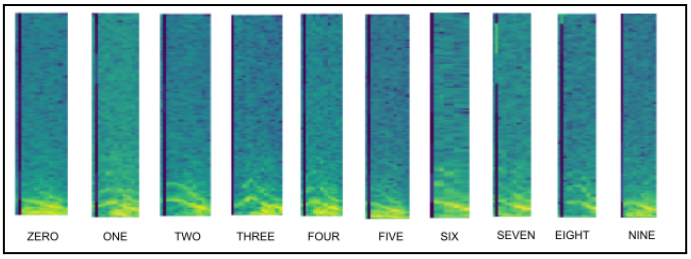

Figure 11: Spectrogram Variations

Spectrograms in ascending order for numbers 0 - 9

4.4 Accuracy

This model utilized 80% data for training and 20% for validation. When classifying the recovered audio files, the training accuracy of the model was 90.11% and the validation accuracy was 97.33% for 10,000 epochs. Figure 12 shows the loss and accuracy graphs for the training and validation datasets.

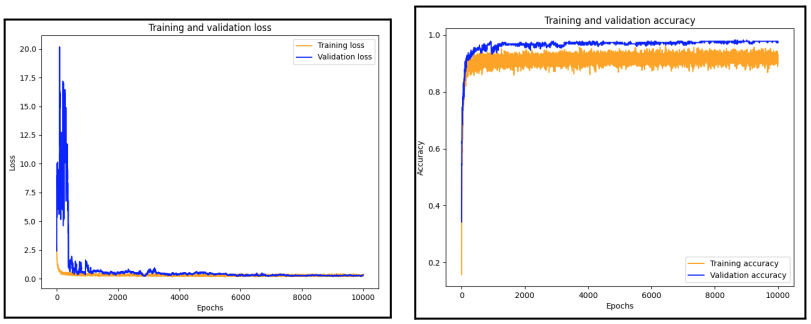

Figure 12: Loss and Accuracy Graphs

Training and Validation Loss and Accuracy Graphs

5 Conclusion and Future Work

This study sought to gather key details and recover sound from tracking and analyzing the visual vibrations. After recovering the audio that resembled the original, this study was able to classify the recovered sound into specific words using a neural network. The recovered sound was imperfect but it was adequate for neural networks to classify. Future plans include enhancing video quality by optimizing the camera, and lighting. Improving the quality of the setup may yield precise data for the dot's movement from the video. Using different surface materials and different objects would allow for the examination of different objects' sensitivities to sound. The plot created from the visual vibrations currently examines movement in the y-axis. Further research may focus on x-axis movements and combinations of planar movements.

5.1 Data and Code Availability

The data, analysis, and code that support the findings of this study are available upon reasonable request.

6 Acknowledgments

I would like to thank Professor Sanjay Ranka for guiding me through the research process and advising me on various aspects of this research project. I am very thankful to my Pioneer personal cohort advisor Ryan Manley for continuous support throughout the process. I would like to express my gratitude to the Pioneer Research Academies institution for connecting me with the professor, for providing access to the Oberlin College digital library, and other tools to guide me through the research process.

References

1. 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)
2. Juhyun Ahn, Yong-Joong Kim and Daijin Kim, "Patch-based visual microphone for improving quality of sound," 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 2016, pp. 3927-3932, doi: 10.1109/ICPR.2016.7900248.
3. M. Hua, L. Zhou, C. Liu and Z. Li, "The Research of Vibration Detection Using the Visual Microphone Technology," 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Changsha, China, 2018, pp. 256-258, doi: 10.1109/ICMTMA.2018.00068.
4. Nassi, B., Pirutin, Y., Shamir, A., Elovici, Y., & Zadov, B. (2020). Lamphone: Real-time passive sound recovery from light bulb vibrations. Cryptology ePrint Archive.
5. Sheinin, M., Chan, D., O'Toole, M., & Narasimhan, S. G. (2022). Dual-shutter optical vibration sensing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16324-16333).
6. Walker, Jearl, et al. Halliday & Resnick's Principles of Physics. tenth ed., Wiley, 2020.
7. Kelm, Robert. "The Nyquist–Shannon Theorem: Understanding Sampled Systems - Technical Articles." All About Circuits, 6 May 2020, www.allaboutcircuits.com/technical-articles/nyquist-shannon-theorem-understanding-sampled-systems/.
8. Christianlillelund. "Classify Mnist Audio Using Spectrograms/Keras CNN." Kaggle, 12 Dec. 2020, www.kaggle.com/code/christianlillelund/classify-mnist-audio-using-spectrograms-keras-cnn.
9. "First Principles of Computer Vision." Fpcv.Cs.Columbia.Edu, fpcv.cs.columbia.edu/. Accessed 6 July 2023.
10. People | MIT CSAIL, people.csail.mit.edu/mrub/papers/RubinsteinPhDThesis.pdf. Accessed 7 July 2023.
11. "What Is a Spectrogram? - Signal Analysis." Vibration Research, 18 May 2023, vibrationresearch.com/blog/what-is-a-spectrogram/.
12. "Free Text to Speech Tool." TextMagic, freetools.textmagic.com/text-to-speech. Accessed 7 July 2023.
13. "Template Matching." OpenCV, docs.opencv.org/3.4/de/da9/tutorial_template_matching.html. Accessed 7 July 2023.
14. Kaehler, Adrian, and Gary R. Bradski. Learning Opencv 3: Computer Vision in C++ with the Opencv Library. O'Reilly, 2017.
15. "How Do Microphones Work?" How Microphones Work, www.mediacollege.com/audio/microphones/how-microphones-work.html. Accessed 8 July 2023.
16. Convolutional Neural Network (CNN) ： Tensorflow Core. TensorFlow. (n.d.). https://www.tensorflow.org/tutorials/images/cnn
17. Visual vibration analysis - Abe Davis. (n.d.). http://www.abedavis.com/thesis.pdf

18. Zhang, Dashan, et al. "A high-speed vision-based sensor for dynamic vibration analysis using fast motion extraction algorithms." *Sensors* 16.4 (2016): 572.
19. Mim, Khatuna Zannat, Abdullah Arafat Miah, and Mohiuddin Ahmad. "Extraction of sound signal from tiny vibrations in motion magnified video using optical flow." 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). IEEE, 2019.
20. Lee, S. Y., et al. "Investigate the impact of colour to grayscale conversion on sound recovery via visual microphone." 2018 2nd International Conference on Imaging, Signal Processing and Communication (ICISPC). IEEE, 2018.
21. Li, Songxu, et al. "mmPhone: Sound Recovery Using Millimeter-Wave Radios With Adaptive Fusion Enhanced Vibration Sensing." IEEE Transactions on Microwave Theory and Techniques 70.8 (2022): 4045-4055.
22. Liu, Ce, et al. "Motion magnification." ACM transactions on graphics (TOG) 24.3 (2005): 519-526.
23. Ahn, Juhyun, and Daijin Kim. "Simple and effective speech enhancement for visual microphone." 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, 2017.
24. Hua, Mingda, et al. "The research of vibration detection using the visual microphone technology." *2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. IEEE, 2018.
25. Chollet, François. *Deep Learning with Python*. Manning Publications, 2021.
26. Ciresan, Dan Claudiu, et al. "Flexible, high performance convolutional neural networks for image classification." *Twenty-second international joint conference on artificial intelligence*. 2011.

Appendix A

A.1 Shannon Nyquist Theorem

The Shannon-Nyquist theorem states that a sampling rate of at least twice the frequency (2N Hz) is required for capturing the information of a specific frequency (N Hz). A sampling rate of 40 kHz is required for capturing the entire range of human hearing which spans from 20 Hz to 20 kHz. Common cameras with the capability of shooting 120 frames per second can only provide an audio frequency of 60 Hz. To include higher frequencies, this study uses a camera that is capable of capturing up to 9,259 frames per second. The highest fps used in this study is only 4800 due to inadequate lighting. Based on the Shannon-Nyquist theorem, only frequencies below 2400 Hz can be captured using this sampling rate and hence any higher frequency found in the reconstructed audio wave would create imperfections in the recovered wave. To eliminate these imperfections, this study uses low-pass filters to eliminate high frequencies. The study also uses a high pass filter at 20 Hz to remove noise in the audio.

A.2 Sound wave propagation

Sound waves are longitudinal - the particles in their path vibrate along the direction of the propagation of the wave. When the sound wave hits a surface, the surface's particles vibrate back and forth in the direction of transmission. Optical cameras in the path of sound waves cannot observe vibrations in objects moving parallel to their lenses since they lack depth perception.

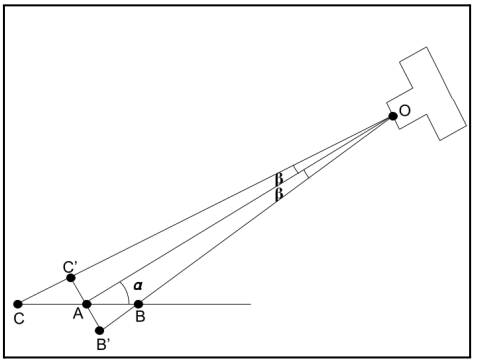

Figure 13: Camera view of dot on paper
Observed movement against actual movement

As shown in Figure 13, let us consider a particle of the surface at point A. This particle oscillates back and forth to points B and C. The camera views the particle at B as B' and the particle at C as C' due to a lack of depth perception. B'AC' is located parallel to the surface of the camera. Below is a proof that the distance captured by the camera sensor is directly proportional to the vibration of the particle.

As ∠AOB' is very small,

$$OB' \approx OA$$

Which implies,

△AOB' is isosceles

From Figure 13, $\angle OAB'$ is 90 =>

$\angle AB'O \approx 90$ degrees

Since $\angle ABB'$ and $\angle OAB$ are alternate interior angles

Let $\angle ABB' = \alpha$

$\sin(\alpha) = AB' / AB$

$AB' = AB \sin(\alpha)$

This proof was integral to this research as it maintains that the camera could perceive the depth by viewing the surface at an angle.

A.3 High-Speed Cameras

Regular videos are taken at 24-to-30 frames per second and are played back at the same rate. A myriad of cameras and phones operate at around 60 frames per second for slow motion. Videos captured at this rate have double the frames as normal videos and playing back these videos at 30 frames per second, allows an individual to see events in slow motion at half the speed. In the slow-motion setting, cameras take a multitude of frames in quick succession.

Frames can be played even slower for visualization of faster movements. This study utilizes this concept to analyze visual vibrations that cannot be perceived by the naked eye. High-speed videos taken in this experiment range from 1000 - 4000 fps. These cameras are commonly used in scientific research, military testing, and industry-wide applications like filming a manufacturing line to tune up the machines and investigating automobile crash testing. In this experiment, the high frame rate allows individuals to see the oscillations that sound creates on the surface.