Improving the Quality of Sound Recovered Using the Visual Microphone with Frame-wise Image Denoising Preprocessing

Ren-Jun Choong\textsuperscript{a}, Wun-She Yap\textsuperscript{b}, Yan Chai Hum\textsuperscript{c} and Yee Kai Tee\textsuperscript{d*}

Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Selangor, Malaysia.

\textsuperscript{a}sakatagintoki@1utar.my, \textsuperscript{b}yapws@utar.edu.my, \textsuperscript{c}humyc@utar.edu.my, \textsuperscript{d}teeyeekai@gmail.com

Abstract. The visual microphone is a passive remote sound recovery technique by analysing subtle vibrations in a video of an object vibrating due to sound. Multiple research endeavours have been done in improving the quality of the recovered sound, but so far none has been done on investigating the effect of denoising the video frame by frame prior to the sound recovery using the visual microphone. This work fills in that gap, and evaluates the quality of the recovered sound based on human intelligibility metrics and simple signal-to-noise ratio metrics. Our best results indicated performing image denoising before sound recovery can reduce the noise power by up to 47.16\%, increase the intelligibility by 9.11\% and signal-to-noise ratio by 17.97\%. It is important to perform image denoising before sound recovery when the recorded videos are very noisy.

1. Introduction

Sound is a form of pressure waves which causes compressions and rarefactions when it propagates through the air. There are various ways to record sound, the most common way is to simply use a conventional microphone. Several methods have also been proposed for recording sound in special environments, mostly for security and surveillance purposes, such as laser-based microphones \cite{1} and using speckles pattern for remote sound extraction \cite{2}.

Knowing that sound waves propagating through the air will induce vibrations on objects which it interacts with, it is shown recently that it is possible to extract sound from a high-speed video footage of the object by analysing those vibrations. The visual microphone (VM) \cite{3} is a name coined to refer to the technique of recovering sound from an object in a silent video by analysing the subtle vibrations. Such subtle vibrations may generate a very subtle visual signal, if it is strong enough to be recorded by a high speed camera, the vibrations can then be amplified using video processing techniques \cite{4} and subsequently be converted to the sound that causes the vibrations.

There are a number of studies done on investigating the effects of manipulating the various aspects of the sound recovery process on the quality of the recovered sound, as well as proposals on improving the quality of the sound recovered using the VM.

A patch-based sound recovery technique is proposed by \cite{5} where the sound is recovered from patches centered about Harris corners \cite{6} in the video. The patch with the best signal-to-noise ratio is selected for sound recovery. The results show improved signal to noise ratio (SNR), as well as the segmental SNR of the recovered sounds but it comes at a cost of decreased intelligibility score.

When the original sound recovery algorithm is proposed, the video motion is calculated by finding
the differences between successive frames with the first frame or first stable frame. By modifying the algorithm slightly [7], approximating the video motion using the difference between successive frames, it is observed that consistent improvements in the quality of the recovered sound are achieved for the log-spectral distance metric.

Later, a speech enhanced visual microphone (SEVM) [8] which made use of nonnegative matrix factorization [9] speech enhancement techniques is proposed. The quality of the sound recovered using the proposed SEVM is reported to be the best under the log likelihood ratio (LLR), the SNR, the segmental SNR, and the cepstral distance measure metrics, when compared with other speech enhancement techniques.

A separate study by [10] proposes a different sound recovery method, which involves breaking the video down into blocks, represented as vectors, and comparing the similarity between vectors from the same block in successive frames with varying amounts of shift. It can be used to determine the object vibrations for sound recovery and reported faster processing time and better sound quality measured using LLR and segmental SNR metrics.

There are also works to investigate which colour to grayscale conversion [11], to simplify sound recovery algorithm, can be used to improve the intelligibility of the recovered sounds. It is reported that a weighted average or a simple average of the red, green and blue colour channels are more suitable for the VM and any visual enhancement of videos would eliminate the useful subtle vibrations thus not recommended.

Furthermore, a study investigates how commonly used image/video compression algorithms used in the closed circuit televisions (CCTVs) can affect the sound recovery [12]. Although the quality of the recovered sound is lower when the higher compression ratio techniques such as H.264 are used, it is still possible to recognize the recovered sound as the input, paving the way to enable outdoor CCTVs for sound recording, despite not having an audio input.

Image denoising is a process to remove the noise in the images. Studies in computer vision and classification tasks have shown that this preprocessing step is important to improve the performance of trained classifier [13]. So far there is still no study to determine whether denoising the individual frames of high-speed footage of objects prior to sound recovery via VM would be useful. In this paper, a few image-based denoising algorithms are investigated to study whether frame-wise denoising preprocessing can enhance the quality of recovered sounds.

2. Methodology

2.1. Video Data

In this work, the experiments are conducted using four videos available online [3]. A reference video frame of each video is provided in Table 1, together with the video parameters. For simplicity and consistency, the videos will be enumerated and referred to by their number for the rest of this text.

| Reference frame | Video 1 | Video 2 | Video 3 | Video 4 |
|-----------------|---------|---------|---------|---------|
| Audio source type | MIDI audio from speaker | MIDI audio from speaker | Human speech | Human speech |
| Video frame rate (Hz) | 2200 | 2200 | 2200 | 20000 |

The four videos in Table 1 are recorded indoors with a Phantom V10 high-speed camera while an audio source is playing beside the object in the video. The objects in the video are illuminated with a photography lamp.
2.2. Image Denoising Algorithms

Each video in Table 1 is experimented with 5 different image denoising algorithms, each with slightly different settings. The image denoising algorithms are first applied to each individual frame of the video, which are subsequently recombined into an output video and saved as a separate file. Hence, each video listed in Table 1 produces total of 16 video files as per the following list. The image denoising algorithms tested are:

1. The N×N average filter, N = 3, 5, 7
2. The Gaussian filter with sigma, σ = 0.50, 0.75, 1.00, 1.25, 1.50, and 1.75
3. The 5×5 guided filter [14], degree of smoothing = 650.25
4. The N×N median filter, N = 3, 5, 7
5. The N×N Wiener filter [15], N = 3, 5, 7

Once the videos have been denoised, they will be used to perform sound recovery via the VM [3].

The VM requires to set the number of scales and orientations to be used in the sound recovery process. For each video file, a range of scales – vary from 1 to 5 and orientations – vary from 1 to 6, are tested, resulting in 30 recovered sound clips per each input video. There are two types of input sound for the 4 videos in Table 1, Musical Instrument Digital Interface (MIDI) audio and human speech. The recovered sounds are processed using different methods, depending on the type of the input sound to further improve the quality of the recovered sound.

For the MIDI sound input videos, a tenth order Butterworth digital bandpass filter is applied with cut-off frequencies at 150 Hz and 400 Hz because the input frequency is within the stated range only. This removes the low frequency noise below 150 Hz, as well as the 120 Hz noise due to the flickering of the photography lamps illuminating the objects while the video is recorded. Whereas, for the human speech input video, a perceptually motivated speech enhancement program based on Bayesian estimators by [16] is applied to improve the quality of the recovered sound. Video 3 is further processed by applying a tenth order Butterworth highpass filter with a cutoff frequency of 150 Hz and a sixth order Butterworth bandstop filter with cutoff frequencies of 295 Hz and 310 Hz.

The filtered/speech enhanced sound files are then evaluated using the short-time objective intelligibility (STOI) metric proposed by [17] and the SNR metric, whereby higher value in each metric represents better quality of sound is recovered.

3. Results and Discussion

Each video produced 510 recovered audio clips ((16 denoising algorithms + 1 original) * 5 scales * 6 orientations per scale) which are compared against the original input sound to determine its quality using both the STOI and SNR metrics. The image denoising algorithm and its corresponding parameters which give the best STOI score for each video is shown in Table 2. It is observed that the recovered sound from Video 3 and 4 had higher STOI than Video 1 and 2. This is because the input sound for Video 1 and 2 is MIDI whereas recorded human speech is used in Video 3 and 4. Since the STOI metric is specifically designed to quantify the intelligibility of human speech, the recovered signal from Video 3 and 4 has higher STOI value than the other two videos, meaning it is easier to be comprehended by humans.

Table 2. The Best STOI value of each video for the different denoising algorithms with different scales, r, and orientations, θ, tested. The higher STOI value refers to higher intelligibility (easier to comprehend).

| Best denoising algorithm (r, θ) | Video 1 | Video 2 | Video 3 | Video 4 |
|-------------------------------|---------|---------|---------|---------|
| Best STOI (denoised)          | 0.708   | 0.706   | 0.611   |         |
| Best STOI (undenoised)        | 0.708   | 0.706   | 0.611   |         |
| Improvement                   | 1.28    | 0.56    | 0.28    | 9.11    |
| Result by [4]                 | 0.379   | 0.239   | 0.612   | 0.447   |
From Table 2, the 5 x 5 average filter is able to perform the best for Video 1, achieving a STOI of 0.541; the 5 x 5 guided filter worked best for Video 2 with the best STOI being 0.311; the guided filter is the best for Video 3, reporting a STOI of 0.708; and the 7 x 7 Wiener filter has the best performance for Video 4, achieving a STOI of 0.611. All our results are better than the results achieved by [4]. It is noted that [4] never actually reported the STOI values of their recovered sound, but they made their recovered sound files available on the internet, and we downloaded them and calculated the STOI ourselves. The STOI scores of the recovered sound signal by [4] are 0.379, 0.239, 0.612, and 0.447 for Video 1, Video 2, Video 3, and Video 4 respectively.

In terms of improvement when compared to the best undenoised videos, the highest improvement is observed in Video 4, which is 9.11 %, followed by Video 1 (1.12 %), Video 2 (0.97 %) and Video 3 (just 0.28 %). There is only marginal improvement in the STOI for the recovered sounds from Video 1 to 3. This is because the Phantom V10 high-speed camera is able to record at 2200 Hz with minimal noise in the recorded videos and the applied tenth order Butterworth digital bandpass filter has removed the uninformative signal out of the input range. The best undenoised results obtained had the same scale and orientation as the best denoised videos, except Video 1 which is achieved by scale of 3 and 1 orientation.

Table 3 shows the image denoising algorithms and its corresponding parameters which give the best SNR score for each video. In Video 1, the 0.5 sigma Gaussian filter produces the best SNR at -5.26 dB. The 1.75 sigma Gaussian filter generates the best SNR in Video 2 at -13.65 dB. In Video 3, the 7 x 7 median filter is the best, where the SNR of the recovered sound is -5.57 dB and in Video 4, the guided filter produces the best SNR value of -7.43 dB. When compared to the undenoised videos, there is only a minor improvement for Video 3 in terms of SNR, which is similar to the STOI study where both achieved < 1% increase. For Video 2, the denoised video recovered 14.47 % better SNR signal. The biggest improvement is obtained in Video 4, the recovered sound from the denoised video is 17.97 % better than the undenoised counterpart. The best undenosed results obtained had the same scale and orientation as the best denoised videos, except Video 2 which is achieved using scale of 2 scales and 2 orientation. Our SNR results are poorer than that of [4], since they use professional audio processing software to improve the SNR of their recovered sound.

Table 3. The Best SNR value of each video for the different denoising algorithms with different scales, r, and orientations, θ, tested. The higher SNR value refers to higher signal-to-noise ratio.

| Best denoising algorithm (r, θ) | Video 1 | Video 2 | Video 3 | Video 4 |
|--------------------------------|---------|---------|---------|---------|
| Best SNR (denoised)            | -5.26 dB| -13.65 dB| -5.57 dB| -7.53 dB|
| Best SNR (undenoised)          | -5.77 dB| -15.96 dB| -5.61 dB| -9.18 dB|
| Improvement                    | 8.84 %  | 14.47 % | 0.71 %  | 17.97 % |

The spectrograms of the best improvements in both SNR and STOI for Video 4 are shown in Figure 1. Since the input speech is only up to 4 kHz, anything above 4 kHz is considered as noise. We calculated the bandpower of the 4 kHz to 10 kHz band. For the undenoised sound recovery, the noise bandpower is 0.000511, it is reduced to 0.000270 when the Wiener filter in Table 2 is used and to 0.000303 when the Guided filter in Table 3 is applied. This indicates that performing image denoising prior to sound recovery can reduce the noise power by 47.16 % and 40.7 % under the STOI and SNR metrics respectively. The percentage of the noise power out of the total power of the signal is also measured, it can be seen that performing image denoising can decrease the percentage of the noise from 11.14 % to below 4.5 %.
An interesting observation is that performing denoising prior to sound recovery will improve the quality of the recovered sound measured in term of STOI and SNR. However, different types of filters are needed to achieve the best STOI and SNR as shown in Table 2 and Table 3. For Video 4, substantial improvement in term of STOI and SNR is obtained after denoising. This is because the video is recorded at 20 kHz. For such a high sampling rate video, the camera shutter exposure time is very short, leading to higher noise in the recorded video. Thus, when the filter is used to minimize the noises by smoothing it, higher quality of sound is recovered.

Substantial improvement in SNR is also obtained for Video 2 because there is a large background in the plant (Table 1). When VM is used to recover the sounds, the whole frame of the video is used to estimate the object vibrations caused by the input sound. Since approximately 50% of the frame of Video 2 consists of the white background which has no informative vibrations for sound recovery but a lot of speckle noises, when an image filter is used to blur out the background noises, a sound clip of higher quality is recovered. The same improvement of Video 2 is not observed for STOI, this may be caused by human comprehension of sound does not have a direct relationship with SNR, meaning a high intelligibility or STOI sound does not necessary mean that it will have high SNR and vice-versa. Other has reported that SNR is not well related to the speech quality [18]. This is also supported by our results here, where all our STOI results are better than the results in [4] whereas it is the reverse for SNR. Thus, it depends on the proposed application of VM, for example, intelligibility would be more important in applications such as a criminal investigation in which the content of the speech is more important. When VM is applied to automatic audio recognition or detection, the SNR would be a more suitable metric for computer processing.

One of the limitations of this work is the limited number of videos tested. It is observed that applying image-based denoising techniques to each frame of a high-speed video is possible to improve the sound recovery using the VM measured in STOI or SNR (Table 2 and Table 3). However, the performance improvement is minimal when the quality of the recorded videos is good such as Video 3. When there is a lot of noise in the recorded videos such as Video 4 due to short shutter exposure time (too high sampling rate), it is important to perform denoising prior to recovery to improve the quality of the recovered sound. Besides that, only a few types of filters are tested, most of them are not able to preserve the features such as edges and borders of the object. Artificial intelligence (AI) based image denoising algorithms which are able to remove the noises in the images without distorting the important object features are not explored.
4. Conclusion and Future Work
Applying image-based denoising techniques to each frame of a high-speed video can improve the sound recovery using the VM measured in both the STOI and SNR. It is important to perform denoising when the recorded videos are very noisy such as recorded using very high sampling rate and less important when the quality of the recorded videos is good. For future work, perhaps AI based image denoising algorithms can be explored because they have been shown to outperform the conventional image-based denoising techniques as tested here, in many computer vision applications.

Acknowledgments
This project is funded by Fundamental Research Grant Scheme (Ref: FRGS/1/2019/ICT04/UTAR/02/1, vote account no: 8073/Y01) and was previously supported by UTAR Research Fund (project number: IPSR/RMC/UTARRF/2016-C2/T04) and Fulbright-MCMC Specialist Grant 2018 (vote account no: 8060/000). The authors would like to thank Abe Davis for sharing the codes and videos used in this work, and the NVIDIA Corporation for the donation of a Quadro P5000 GPU.

References
[1] Takács G, Otčenáš J, Vachálek J and Rohal’-Ilkiv B 2016 Modal response-based technical countersurveillance measure against laser microphones J. Vibroengineering 18 3369–82
[2] Chen Z, Wang C, Huang C, Fu H, Luo H and Wang H 2014 Audio signal reconstruction based on adaptively selected seed points from laser speckle images Opt. Commun. 331 6–13
[3] Davis A, Rubinstein M, Wadhwa N, Mysore G J, Durand F and Freeman W T 2014 The visual microphone: Passive recovery of sound from video ACM Trans. Graph. 33
[4] Wu H Y, Rubinstein M, Shih E, Guttag J, Durand F and Freeman W 2012 Eulerian video magnification for revealing subtle changes in the world ACM Trans. Graph. 31 1–8
[5] Ahn J, Kim Y J and Kim D 2016 Patch-based visual microphone for improving quality of sound Proc. - Int. Conf. Pattern Recognit. 0 3927–32
[6] Harris C and Stephens M 2013 A Combined Corner and Edge Detector Alvey vision conference vol 50 TS-E, ed C G Harris and M Stephens (Citeseer) pp 23.1-23.6
[7] Yasumi Y, Takiguchi T and Ariki Y 2017 Visual Sound Recovery Using Momentary Phase Variations Int. Work. Front. Comput. Vis. 2–5
[8] Ahn J and Kim D 2018 Simple and effective speech enhancement for visual microphone Proc. - 4th Asian Conf. Pattern Recognition, ACPR 2017 700–5
[9] Mohammadiha N, Smaragdis P and Leijon A 2013 Supervised and unsupervised speech enhancement using nonnegative matrix factorization IEEE Trans. Audio, Speech Lang. Process. 21 2140–51
[10] Shabani M A, Samadfam L and Sadeghi M A 2017 Local visual microphones: Improved sound extraction from silent video Br. Mach. Vis. Conf. 2017, BMVC 2017
[11] Lee S Y, Yap W S, Hum Y C, Goi B M and Tee Y K 2018 Investigate the Impact of Colour to Grayscale Conversion on Sound Recovery via Visual Microphone 2018 2nd Int. Conf. Imaging, Signal Processing. Commun. ICISP 2018 138–42
[12] Lee S Y, Yap W S, Hum Y C, Kwan B H, Goi B M, Tee Y K and Jin Z 2018 Study the effect of commonly used video compression techniques on sound recovery via negligible object vibrations for visual surveillance system ed Unknown ACM Int. Conf. Proceeding Ser. 111–5
[13] Katkar V, Bhatia D and Kulkarni S 2015 Content based video classification using fuzzy rule base for edge detection 2015 Int. Conf. Nasent Technol. Eng. Field, ICNTE 2015 - Proc. 1-6 TS-RIS M4-Citavi
[14] He K, Sun J and Tang X 2013 Guided image filtering IEEE Trans. Pattern Anal. Mach. Intell. 35 1397–409
[15] Lim J S Two-dimensional signal and image processing ph
[16] Loizou P C 2005 Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum *IEEE Trans. Speech Audio Process.* **13** 857–69

[17] Taal C H, Hendriks R C, Heusdens R and Jensen J 2011 An algorithm for intelligibility prediction of time-frequency weighted noisy speech *IEEE Trans. Audio, Speech Lang. Process.* **19** 2125–36

[18] Kondo K 2012 Speech Quality ed K Kondo (Berlin, Heidelberg: Springer Berlin Heidelberg) pp 7–20