Audio Signal Extraction and Enhancement Based on CNN From Laser Speckles

Chang Liu, Lin Li, Xueyuan Huang, Xiaozhong Wang, and Cheng Wang, *Senior Member, IEEE*

***Abstract*—A micro vibration signal extraction method based on deep neural network is proposed. Rough surface of vibrating object modulates the illuminating laser wave front and generates speckle pattern, which is recorded by a linear array CMOS and preprocessed and input into a 16-layer convolution neural network (CNN) trained with specially prepared data. The optical experimental setup is analyzed to fulfil the temporal and spatial Shannon sampling theorem. The output audio signals are evaluated with standard algorithms and show enhanced segmental SNR and intelligibility. The effects of different input audio types and quality of raw audio signals are investigated, and the results show that the neural network is robust to the input. The CNN structure is optimized and the results show the performance decrease with the reduction of convolution layers. The performances of three popular deep neural networks are compared and the performance of CNN is better.**

***Index Terms*—Speckle, audio signal extraction and enhancement, convolution neural network.**

I. Introduction

WHEN a coherent laser beam is used to illuminate a rough surface, the scattered light will generate a random light field, which is known as the secondary speckle [1]. The features of surface and its substrate are embodied in the speckle patterns, which can be used in flood flow imaging [2], vibration detection [3], material identification [4], and many others [5]. In 2009, Z. Zalevsky extracted audio and heart beat signal through image correlation algorithm from speckle pattern sequences recorded by a high-speed camera [3]. Compared to the laser Doppler vibration detection method [6], the speckle imaging method has the advantages of long detection distance, simple system setup, loose surface requirement of vibrating object. According to the Nyquist-Shannon sampling theorem, the frame rate of the camera should be at least twice the highest vibration frequency of target, which challenges the imaging frame rate of the camera used in some applications.

Manuscript received November 9, 2021; revised December 6, 2021; accepted December 16, 2021. Date of publication December 21, 2021; date of current version January 4, 2022. This work was supported by the Natural National Science Foundation of China under Grant 61975167. *(Corresponding author: Xiaozhong Wang.)*

Chang Liu, Lin Li, Xueyuan Huang, and Xiaozhong Wang are with the Department of Electronics Engineering, School of Electronics Science and Engineering, Xiamen University, Fujian 365001, China (e-mail: 253697168@qq.com; 23120191150220@stu.xmu.edu.cn; 23120201150227@stu.xmu.edu.cn; wangxz@xmu.edu.cn).

Cheng Wang is with the Department of Computer Science, Xiamen University, Xiamen 361005, China (e-mail: cwang@xmu.edu.cn).

Digital Object Identifier 10.1109/JPHOT.2021.3136908

To overcome the problem of frame rate, Veber *et al.* exploited a commercially available photodiode to extract vibration signals from photocurrent generated by the light flux of speckle [7]. A mask is put between the imaging lens and the detector to enhance the sensitivity. Furthermore, Bianchi optimized the beam size of the probing laser and the detector's aperture to suppress the distortion and enhance the SNR and intelligibility of the recovered audio signal [8]. In order to improve the SNR and decrease the data flow, Bianchi *et al.* exploited a linear-scan CCD to record the speckle patterns [9]. Intensity correlation method is used to extract the audio signal and the maximum detection distance of 300 m is realized. To enhance the signal processing speed, the gray value of recorded speckle pattern is used to recover the audio signal. The calculation time is reduced and the SNR of the reconstructed audio signal is increased [10]. Recently, optical flow algorithm is introduced to laser speckle analysis and real-time audio detection is realized [11]. In our previous work, the multi-channel signals of a linear CMOS array are fused, and the SNR and intelligibility of the recovered audio signals are enhanced [12].

To extract audio signals from optical speckle patterns, the principle signal processing algorithms include cross-correlation, flux variations, optical flow, channel fusion, complex pyramid [13], and machine learning. Barcellona *et al.* compared the commonly used algorithms and concluded that the machine learning method obtained the best performance [14]. However, detailed investigation of machine learning method is not introduced in that paper. To our knowledge, there is no detailed report of machine learning method used to extract audio signal from speckle patterns up to now. In contrast, machine learning method is used in various fields and excellent performances are obtained [15], [16]. Therefore, investigation of audio signal extraction and enhancement using machine learning method in speckle vibration detection field is necessary and interesting.

In this paper, a convolutional neural network (CNN) is specially trained to extract and enhance audio signal from speckle patterns, which is generated by illuminating a 532 nm laser beam on a vibrating rough surface. The speckle is collected by a photograph lens and recorded by a linear CMOS array. The CNN training sets are obtained by randomly mixed noisy audio segments in our experiment with pure audio samples from TIMIT speech dataset. The CNN structure is optimized for audio signal extraction and enhancement. The recovered audio signals are evaluated by standard audio signal evaluating algorithms, including segmental SNR (SegSNR) [17], log likelihood ratio (LLR) [18] and normalized subband envelope correlation

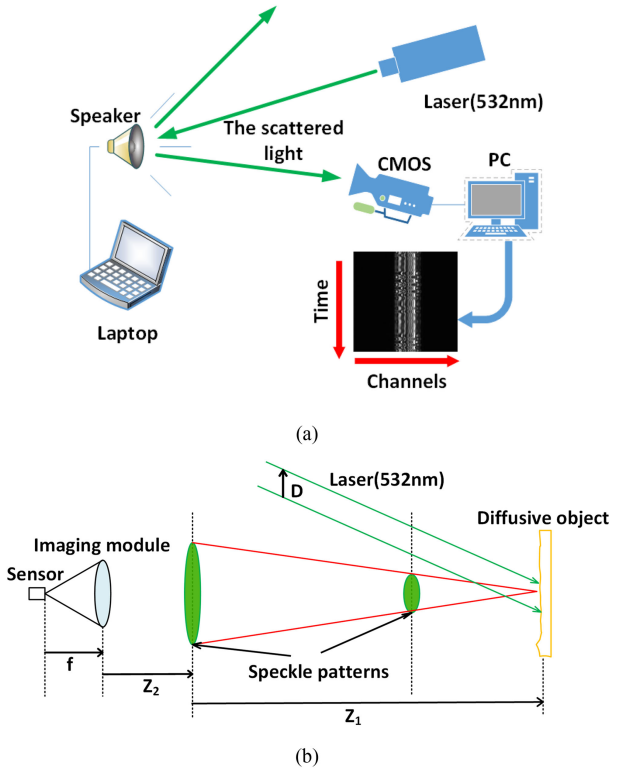

Fig. 1. (a) Experimental setup. The inset shows a picture of typical recorded data. The vertical axis of the inset represents time, and the horizontal axis shows the gray values of the linear CMOS. (b) Schematic description of the system.

(NSEC) [19]. The results show that CNN can enhance the SegSNR and intelligibility of the extracted audio signals.

In addition, speech enhancement based on recurrent neural network (RNN) with gated recurrent unit (GRU) [20] and variance constrained autoencoder (VCAE) [21] are also used to compare the performances of different deep neutral networks. The rest of the paper is organized as follows. The experimental setup and CNN structure used are introduced in Section 2. The signal processing models, results and analysis are presented in Section 3. Section 4 is a brief conclusion.

II. Experimental Setup and CNN Structures

The experimental setup is shown in Fig. 1(a). Beam from a 532 nm green laser (CNI laser MSL-III-532-50 mW) is collimated and illuminated on the diaphragm of an ordinary table speaker, which is droved by a laptop computer with different kinds of audio signals in Chinese and English. The vibrating diaphragm is used as the micro-vibration excitation source. The surface of the diaphragm is rough enough to generate speckles, which is collected via a photography imaging lens (Sigma Zoom Master with focal length 35 mm∼70 mm) and recorded by a linear array CMOS (Basler racer raL2048-48gm). Tilt of the vibrating diaphragm generates a transverse shift of the speckle pattern, which can be recorded by a defocused imaging lens and used to recover the source vibrating signal.

According to the Nyquist-Shannon sampling theorem, the experiment needs to satisfy the following two conditions. Firstly, the frame rate of the camera should be at least twice the highest vibration frequency of the signal. The audio signal usually can be represented in the frequency range of 300-3400 Hz. Therefore, the sampling line rate is set to 8000 Hz in our experiment. Secondly, the speckle size should be greater than two pixels in length. Speckle can be divided into objective speckle and subjective speckle. Objective speckle is the speckle formed through free space transmission, and subjective speckle is the speckle formed through the imaging system. The relationship between objective speckle size (S_o) and subjective speckle size (S_s) is approximately described as (1).

$$\frac{S_o}{S_s} = \frac{Z_2}{f} \tag{1}$$

where Z_2 is the distance between the objective speckle and the imaging system (the object distance of the imaging system) and f is the focal length of the imaging system. Schematic diagram of the experimental system is shown in Fig. 1(b). The mean size of objective speckle can be expressed as,

$$S_o = 1.22\frac{\lambda Z_1}{D} \tag{2}$$

where λ is the wavelength of laser, Z_1 is the distance from the rough surface to the objective speckle and D represents the diameter of illuminating laser spot. According to (1) and (2), Ss may be expressed as follows:

$$S_s = 1.22\lambda\frac{Z_1 f}{Z_2 D} \tag{3}$$

The appropriate subjective speckle size can be obtained by controlling these parameters.

Fig. 2 shows the structures of the CNN network and the feature loss network. The CNN used in this experiment is composed of 16 convolution layers [22]. The first layer and the last layer represent the noisy input speech signal and the enhanced output speech signal, respectively. Both layers are one-dimensional vectors. The middle layers (2-15) are two-dimensional n $\times$ W tensor, where W is the number of feature maps in each layer, which is set to 64 in this network. The content of each layer is calculated by 3 $\times$ 1 convolution check on the data of the previous layer, followed by an adaptive normalization and a pointwise leaky rectified linear unit (LReLU) activation. The parameters of adaptive normalization are determined by back propagation method.

The loss function is based on a feature loss network, which is consisted of a stack of 15 convolutional layers. Each feature layer (i.e., layers 2 through 15) is computed from the previous layer via a convolution with 3$\times$1 learned kernel, followed by batch normalization and a pointwise LReLU. Layer 1 corresponds to the input signal with length N. Other layers are half the length of the previous layer. The width of the first layer is 1. The widths of the other layers (2-14) are $32\times2^{[(L-2)/5]}$, where L is the layer number. The length of the 15th layer, which is also the output layer and obtained through average-pooling, is 1. The loss function is based on the difference between output of the feature loss network and the output of the denoising network being trained.

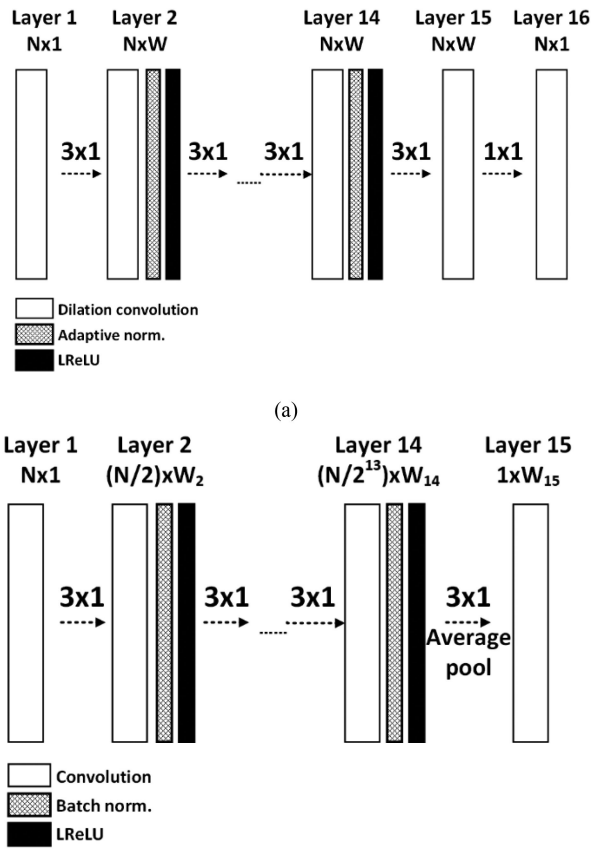

Fig. 2. (a) The structures of the CNN network, (b) the structure of the feature loss network.

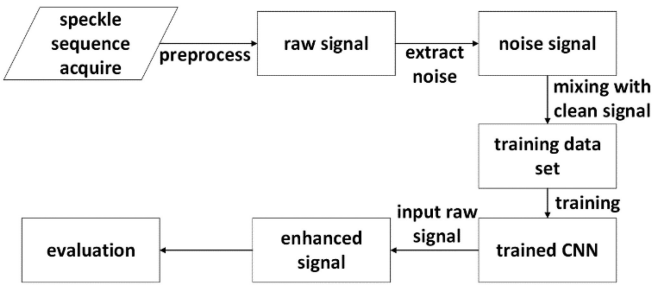

Fig. 3. Flow chart of the signal processing.

III. Signal Process and Analysis

A. Signal Process

Signal processing flow chart is shown in Fig. 3. Speckle pattern data recorded by a linear array CMOS are preprocessed and used as raw audio signals, from which noise segments are selected. Through randomly mixing these noise signals with pure audio signals from TIMIT speech dataset, specialized training set is prepared. This training set is used to train the adopted CNN. Consequently, a trained network specialized to our application is obtained. The raw audio files are input into the trained network and the output enhanced signals are evaluated and analyzed.

The recorded speckle patterns should be transformed into the input data format of the CNN. Firstly, the CMOS recorded data is converted into a matrix, the gray values of each column corresponding to a vibration signal. Then, the gray values of each pixel in that column are squared and summed. The sum represents the energy of the respective vibration signal. Ten strongest signals are selected as seed signals. Finally, every seed signal is processed to eliminate the DC component and normalized to fulfil the 16k sample requirement of the CNN used. These signals are written into a .wav file respectively using the audiowrite function in MATLAB and are the raw audio signals, which are used as input for the CNN.

In order to enhance the performance of the network, specialized training sets are prepared. As the play and record of the audio signals are asynchronous, there are noise parts at both ends of the raw audio signal. 100 non-repeating noise segments are intercepted from the raw audio files using Adobe Audition software. Significantly, this noise signal is specified to our experimental environment. By randomly mixing these 100 noise signals with 6200 pure audio samples downloaded from TIMIT speech dataset, 6200 groups of training data are obtained. Consequently, the noise mixed audio signals and the corresponding clean audio signals are aligned in time, which is important for the training. Among the 6200 groups of training data, 5000 groups are used for the training and 1200 groups are used for testing. The network is trained on a server with a GeForce RTX 3070 GPU and 10GB video memory and lasted for 1000 epochs.

In the first experiment, a piece of 4 seconds length audio from TIMIT dataset is used as input signal and the measuring distance is about 5 m. The recorded speckle patterns are transformed into raw file and used as input for the trained 16 layers CNN. The results are shown in Fig. 4. The waveform of the original, the raw input of CNN and the CNN enhanced audio signal are shown in the left column from up to bottom. We can see that the noise and distortion are both decreased in the CNN enhanced signal. The respective spectrograms are shown in the right column of Fig. 4. We can see that the amplitudes are increased in the enhanced audio signals and the increase in the high frequency component is more evident.

B. Effect of Audio Types and Quality

The enhanced audio signals are evaluated by SegSNR, LLR and NSEC algorithms and the results are shown in Table I. From Table I we can see that the SegSNR score is enhanced by 4.9 dB (from -14.48 to -9.58) for the single English audio signal. Ideally, the maximum LLR and NSEC are 0 and 1, respectively. LLR evaluates the likelihood between two signals. NSEC evaluates the intelligence of the audio signal. For the single English signal, the LLR and NSEC are enhanced by 8.9% and 13.8%, respectively. From Table I we can see that the CNN can also enhance the Chinese speech and song audio signals. For Chinese audio signal, the SegSNR, LLR and NSEC are enhanced by 53.8%, 12.1% and 22.4% respectively. For the song, the SegSNR and LLR are enhanced by 71.3% and 10.8% respectively. The NSEC evaluation results for the song

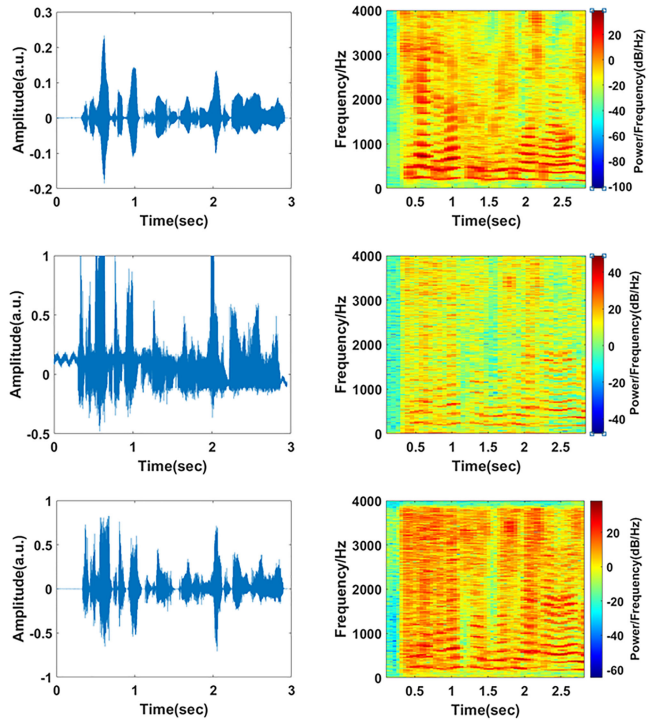

Fig. 4. Waveforms and spectrograms of the input and recovered audio signals. The left column shows the waveforms of original input, raw signal and CNN enhanced audio signal from up to down. Right column is their respective spectrograms.

TABLE I
TYPICAL EVALUATION RESULTS FOR DIFFERENT AUDIO SOURCE TYPES

Audio type	method	SegSNR	LLR	NSEC
English	Raw	-14.48	1.57	0.58
	CNN	-9.58	1.43	0.66
Chinese	Raw	-15.64	1.65	0.58
	CNN	-7.23	1.45	0.71
song	Raw	-9.30	1.57	0.06
	CNN	-2.67	1.40	0.05

are bad. The authors think that the main reason is that the NSEC algorithm is not suitable to evaluate song and may consider the music background as noise.

Either in experiments or in future application, laser vibration measurement must face the problem of acquisition quality. The power fluctuations of the laser, the adjustment of the collecting system, can both effect the quality of the recorded signal. The effect of raw audio quality on the performance of CNN are shown in Table II. In this paper, we classify the audio quality according to their SegSNR evaluation results into low (smaller than -18dB), middle (-18dB $\sim$ -15dB) and high (larger than -15dB), respectively. From Table II we can see that the SegSNR are increased by 19.0%, 34.8% and 33.8% for low-, middle- and high-quality audio signals, respectively. Accordingly, the LLR are increased by 2.4%, 8.3% and 10.3%, respectively. The NSEC are enhanced

TABLE II
EFFECT OF QUALITY OF RAW AUDIO SIGNAL

Audio quality	method	SegSNR	LLR	NSEC
low	Raw	-18.34	1.57	0.43
	CNN	-14.85	1.51	0.53
middle	Raw	-17.78	1.56	0.44
	CNN	-11.60	1.43	0.54
high	Raw	-14.48	1.45	0.58
	CNN	-9.58	1.30	0.66

TABLE III
TYPICAL EVALUATION RESULTS FOR DIFFERENT LAYERS

	SegSNR	LLR	NSEC
none	-14.48	1.56	0.44
CNN-14	-9.58	1.43	0.66
CNN-10	-13.42	1.52	0.62
CNN-6	-13.87	1.68	0.50
CNN-2	-14.15	1.61	0.46

TABLE IV
TYPICAL EVALUATION RESULTS FOR DIFFERENT NETWORKS

	SegSNR	LLR	NSEC
raw	-14.48	1.57	0.58
CNN	-9.58	1.43	0.66
GRU	-12.03	1.42	0.63
VCAE	-11.15	1.54	0.65

by 23.3%, 22.7% and 13.8%, respectively. The results show that the enhancement of SegSNR and LLR for low-quality audio signal is not as good as that for middle- and high-quality audio signals. These results show that an optimized signal collecting system is important for obtaining high performance.

C. *Effect of Network Structure*

In order to evaluate the effect of network structure on the performance, CNN with 14, 10, 6 and 2 convolution layers are trained utilizing the same training set. The output audio signals are evaluated and the results are shown in Table III. With the reduction of convolution layers, the performance degrades quickly. The authors think the main reason is that the mapping from the input to the output is not complete with reduced convolution layers.

The performance of different type of deep neural networks are compared and the results are shown in Table IV. The source audio signal is 4 seconds length English audio. The performance of the 16-layers CNN is used in the comparison. The RNN includes a total of 215 units, 4 hidden layers, with the largest layer of 96 units [20]. Three hidden layers are realized with gated

recurrent unit (GRU), which make it suitable for training with small samples. VCAE has an encoder/decoder structure with a learned latent representation, which prevents overfitting and improves generalization [21]. VCAE is selected as a reprehensive of unsupervised learning network.

From the viewpoint of SegSNR, the performance of CNN is best, that of VCAE is in the middle. The performance of GRU is worst. From the viewpoint of LLR, the performance of CNN and GRU are almost the same and both are better than that of VCAE. The NSEC results show that the performance of CNN and VCAE are almost the same and both are a little better than that of GRU. Therefore, the performance of CNN is best in the deep neural networks used in this paper.

The different performances of the exploited neutral networks may partly be attributed to the fact that only CNN is specially trained exploiting the costume prepared training data set. The others are trained using general data sets and applied directly in this manuscript. The second important reason is that the number of trainable variables are different for these networks. The numbers of trainable variables of the exploited networks are 160k, 87k, 948k for CNN, GRU and VCAE, respectively. The VCAE owns the maximum number of trainable variables. However, its performance is inferior to that of CNN. Therefore, the authors think that VCAE is not suitable for the process of our audio signals. Generally, GRU is appropriate for the process of audio signals. In this paper, its performance is inferior to that of CNN. The main reason may partly be that its number of trainable variables is about half that of CNN.

It is found in experiments that changes of the speaker surface or the speaker device would affect the performance of the trained network. The authors think that the main reason is that the detailed noise feature changes with the variation of speaker surface or the speaker device. To improve the generalizability of the trained neural network, noise signals extracting from experiments with different speaker surfaces or speaker devices should be used in the preparation of the training data set.

IV. Conclusion

In this paper, a micro vibration signal extraction method based on deep neural network is proposed. Laser speckle patterns generated by the surface scattering of the vibrating object is recorded by a linear CMOS camera and preprocessed and used as input for the trained CNN. The 16-layer CNN is specially trained utilizing custom dataset, which is prepared by randomly mixing the noise segment in our raw audio files with pure audio signals from TIMIT dataset. The output of CNN is evaluated using SegSNR, LLR and NSEC algorithms.

The effects of audio types, raw audio quality and network structures on performance are investigated. The results show that audio types cannot affect the quality of the processing results evidently, which means the neural network is robust to the input audio type. The performance for low quality input raw file is bad. On the contrary, the performance for middle- and high-quality input raw files are almost same and good. These results show that an optimized signal collecting system is important for obtaining high performance. The performances of three popular deep neural networks, CNN, GRU and VCAE, are compared and the comprehensive performance of CNN is slightly better, which is attributed to the specialized training set on the one hand and the appropriate number of trainable variables on the other hand.

Acknowledgment

The authors thank Dr. Qingyang Hong of the Department of Artificial Intelligence, Xiamen University for fruitful discussions of this work.

References

[1] J. W. Goodman, *Speckle Phenomena in Optics: Theory and Applications*. Englewood, Colorado, USA: Roberts & Company, 2007.
[2] S. Janaka, R. Abhishek, L. Nan, and T. Nitish V, "Laser speckle contrast imaging: Theory, instrumentation and applications," *IEEE. Trans. Biomed. Eng.*, vol. 6, pp. 99–110, Jan. 2013.
[3] Z. Zalevsky, Y. Beiderman, and I. Margalit, "Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern," *Opt. Exp.*, vol. 17, no. 24, pp. 21566–21580, Nov. 2009.
[4] A. M. Stolyarov, R. M. Sullenberger, D. R. Crompton, T. H. Jeys, B. G. Saar, and W. D. Herzog, "Photothermal speckle modulation for noncontact materials characterization," *Opt. Lett.*, vol. 40, no. 24, pp. 5786–5789, Dec. 2015.
[5] R. S. Sirohi, *Speckle Metrology*. New York, NY, USA: Marcel Dekker, 1993.
[6] R. Li, T. Wang, Z. Zhu, and W. Xiao, "Vibration characteristics of various surfaces using an LDV for long-range voice acquisition," *IEEE Sens. J.*, vol. 11, no. 6, pp. 1415–1422, Jun. 2011.
[7] A. A. Veber, A. Lyashedko, and E. Sholokhov, "Laser vibrometry based on analysis of the speckle pattern from a remote object," *Appl. Phys.*, vol. 105, no. 3, pp. 613–617, Nov. 2011.
[8] B. Silvio, "Vibration detection by observation of speckle patterns," *Appl. Opt.*, vol. 53, no. 5, pp. 931–936, Feb. 2014.
[9] S. Bianchi and E. Giacomozzi, "Long-range detection of acoustic vibrations by speckle tracking," *Appl. Opt.*, vol. 58, no. 28, pp. 7805–7809, Oct. 2019.
[10] Z. Chen, C. Wang, and C. Huang, "Audio signal reconstruction based on adaptively selected seed points from laser speckle images," *Opt. Commun.*, vol. 331, pp. 6–13, Nov. 2014.
[11] N. Wu and S. Haruyama, "Real-time audio detection and regeneration of moving sound source based on optical flow algorithm of laser speckle images," *Opt. Exp.*, vol. 28, no. 4, pp. 4475–4448, Feb. 2020.
[12] C. Dai, C. Liu, Y. Wu, X. Wang, H. Fu, and H. Sun, "Audio signal detection and enhancement based on linear CMOS array and multi-channel data fusion," *IEEE Access*, vol. 8, pp. 133463–133469, Jul. 2020.
[13] A. Davis, M. Rubinstein, and N. Wadhwa, "The visual microphone: Passive recovery of sound from video," *ACM Trans. Graph*, vol. 33, no. 4, Jul. 2014.
[14] C. Barcellona *et al.*, "Remote recovery of audio signals from videos of optical speckle patterns: A comparative study of signal recovery algorithms," *Opt. Exp.*, vol. 28, no. 6, pp. 8716–8723, Mar. 2020.
[15] G. E. Karniadakis, I. G. Kevrekidis, and L. Lu, "Physics-informed machine learning," *Nat. Rev. Phys.*, vol. 3, pp. 422–440, May 2021.
[16] O. Avci, O. Abdeljaber, S. Kiranyaz, M. Hussein, M. Gabbouj, and D. J. Inman, "A review of vibration-based damage detection in civil structures: From traditional methods to machine learning and deep learning applications," *Mech. Syst. Signal Process.*, vol. 147, Jan. 2021, doi: 10.1016/j.ymssp.2020.107077.
[17] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, pp. 2819–2822.
[18] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Signal Process.*, vol. SP-24, no. 5, pp. 380–391, Oct. 1976.
[19] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proc. EUSIPCO*, 2009, pp. 1849–1853.
[20] J. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *Proc. MMSP*, Vancouver, Canada, 2018, pp. 1–5.
[21] D. T. Braithwaite and W. B. Kleijn, "Speech enhancement with variance constrained autoencoders," in *Proc. INTERSPEECH*, 2019, pp. 15–19.
[22] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," 2018, *arXiv: 1806.10522.*