Note: Sound recovery from video using SVD-based information extraction

Dashan Zhang, Jie Guo, Xiujun Lei, and Chang'an Zhu

Citation: Rev. Sci. Instrum. **87**, 086111 (2016); doi: 10.1063/1.4961979
View online: http://dx.doi.org/10.1063/1.4961979
View Table of Contents: http://aip.scitation.org/toc/rsi/87/8
Published by the American Institute of Physics

Note: Sound recovery from video using SVD-based information extraction

Dashan Zhang, Jie Guo,[a] Xiujun Lei, and Chang'an Zhu
Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China

(Received 6 July 2016; accepted 18 August 2016; published online 30 August 2016)

This note reports an efficient singular value decomposition (SVD)-based vibration extraction approach that recovers sound information in silent high-speed video. A high-speed camera of which frame rates are in the range of 2 kHz–10 kHz is applied to film the vibrating objects. Sub-images cut from video frames are transformed into column vectors and then reconstructed to a new matrix. The SVD of the new matrix produces orthonormal image bases (OIBs) and image projections onto specific OIB can be recovered as understandable acoustical signals. Standard frequencies of 256 Hz and 512 Hz tuning forks are extracted offline from their vibrating surfaces and a 3.35 s speech signal is recovered online from a piece of paper that is stimulated by sound waves within 1 min. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4961979]

The ability of dynamic extraction of remote sounds is very appealing and has been used before mainly for surveillance and security. Existing laser-based active methods for remote sound collecting have strict retro-reflective or specular requirements on vibrating surface.[1–3] Moreover, laser Doppler interferometry has been widely applied on the measurement of aero-acoustic phenomena.[4,5] As an emerging technology, image-based sound recovery techniques extract vibration signals from filmed video files by using high-speed camera systems. Without relying on additional sensors or detection modules, phase-based algorithms have been applied to magnify and visualize extremely subtle motions.[6] Speech information and material properties can be recovered and estimated from the spatial vibrations in video frames.[7–9] However, subjecting to the complexity of computation, phase-based approaches[7] and DIC techniques[10] are commonly applied in post-processing.

In this note, we demonstrate an efficient approach that recovers sound from video based on singular value decomposition (SVD). The procedure of the offline passive SVD-based vibration recovery is shown in Fig. 1. Given a video containing n frames, sub-images $R_i(i = 1,2,\ldots,n)$ of which size is $h \times w$ are selected in each frame at same location in image coordinates. These sub-images are transformed into column vectors and are combined into a new matrix, which is described as $\mathbf{G}_{n\times(w\times h)} = [R_1(:), R_2(:), \ldots, R_n(:)]^T$. The magnitudes of the singular values represent the energy distribution of expansion in the decomposition; thus, the former k terms can be truncated as the approximation of the original matrix in the SVD decomposition,

$$\begin{aligned}\mathbf{G}_{n\times(w\times h)} &\approx \mathbf{U}_{n\times k}\cdot\mathbf{D}_{k\times k}\cdot\mathbf{V}^T_{(w\times h)\times k}\\ &\approx \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \cdots + \sigma_k\mathbf{u}_k\mathbf{v}_k^T, \qquad (1)\end{aligned}$$

where $\sigma_i(i = 1,2,\ldots,k)$ are the singular values listed in decreasing order, and the left-singular vectors $\mathbf{u}_i \in \mathbb{R}^{n\times 1}$, $(i = 1,2,\ldots,k)$ and the right-singular vectors $\mathbf{v}_i \in \mathbb{R}^{(w\times h)\times 1}$, $(i = 1,2,\ldots,k)$ are both orthonormal to one another in each vector group.

From this equation, we can see that the right-singular vectors form orthonormal bases of $w \times h$-dimensional space. Given the length of the right-singular vectors are all $w \times h$, $\mathbf{v}_i$ can be rewritten as matrix form as $I_{i,w\times h}(i = 1,2,\ldots,k)$. Thus, the right-singular vectors of matrix $\mathbf{G}_{n\times(w\times h)}$ form a set of orthonormal image bases (OIBs). Corresponding to the ith component, vector $\sigma_i\mathbf{u}_i$ is a coordinate of the projection of matrix $\mathbf{G}_{n\times(w\times h)}$ on image basis vector $\mathbf{v}_i$ (or $I_{i,w\times h}$),

$$\begin{aligned}\mathbf{G}\cdot\mathbf{v}_i &= [R_1(:), R_2(:), \ldots, R_n(:)]^T\cdot\mathbf{v}_i\\ &= [R_1\otimes I_{i,w\times h}, R_2\otimes I_{i,w\times h}, \ldots, R_n\otimes I_{i,w\times h}]^T\\ &= [\sigma_1 u_{i1}, \sigma_2 u_{i2}, \ldots, \sigma_n u_{in}]^T, \qquad (2)\end{aligned}$$

where the operation $\otimes$ is defined as follows: given two matrices ($\mathbf{A}$ and $\mathbf{B}$) with the same dimension, $\mathbf{A}\otimes\mathbf{B} = \sum_i\sum_j a_{ij}b_{ij}$. Thus, $\sigma_i\mathbf{u}_i$ is the projection signal of these n sub-images on ith orthonormal image basis, which reveals the vibration information of the video in the ith principal direction.

To validate the effectiveness of the proposed offline passive vibration recovery approach, a confirmatory experiment was conducted to recover vibration frequencies of tuning forks. As shown in Fig. 2(a), a high-speed camera of which frame rates are in the range of 2 kHz–10 kHz was used to film the vibration of two aluminum alloy tuning forks. The standard frequencies of these two tuning forks are 256 Hz and 512 Hz. After setting vibrating by striking these tuning forks against a surface with a small excitation, we filmed the vibration over 1 s after waiting a moment to allow high overtones to die out. The image size of the captured videos is 320×80 pixels and the size of both the selected sub-images is 50×50 pixels as shown in Fig. 2(b).

Fig. 3 gives the vibration extraction result of these two tuning forks recovered from 3 kHz videos. For the reason that the vibration of sound sources (tuning forks) was clear and the noise of the environment influenced little on rigid aluminum alloy, the vibration information of both of these tuning forks was found in their first component signals. The root-mean-square (RMS) envelops (use a window with a length of

[a]Author to whom correspondence should be addressed. Electronic mail: guojiegj@mail.ustc.edu.cn

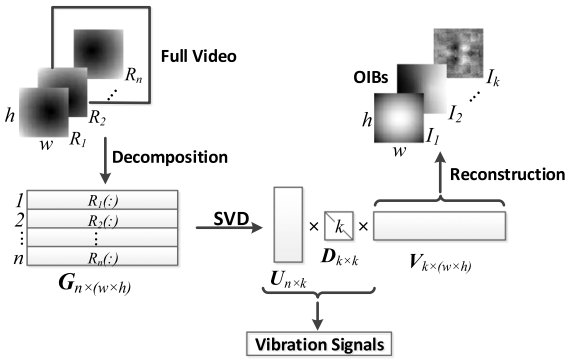

FIG. 1. Procedure of the offline vibration recovery.

50 samples) marked in colored lines reveal the damped vibration of the recovered signals and the spectrograms match accurately to the standard frequencies of these two tuning forks.

Vision-based sound recovery approaches often suffer from low efficiency and cannot return real-time speech signal.[7,10] However, by slightly modifying the proposed offline approach, this sound recovery procedure becomes online and highly efficient. The schematic of the online sound recovery is shown in Fig. 4. Since the acquisition of OIBs of the captured video is flexible, the OIBs can be obtained by using a small number of frames at the beginning of the video. Given a video containing n frames, useable OIBs can be calculated and selected by using sub-images in preceding m ($m < n$) frames. For the upcoming $m + 1$ frame, vibration signal is obtained by projecting the sub-image in $m + 1$ frame on the specific OIB,

$$C_{m+1} = R_{m+1} \otimes I_s, \tag{3}$$

where C_{m+1} refers to the calculated vibration signal in the $m + 1$ frame; R_{m+1} is the selected sub-image; and I_s is the specific OIB for effective projection. The obtained vector $\mathbf{C}_s = [C_{m+1}, C_{m+2}, \ldots, C_n]$ is considered to be the vibration signal corresponding to the specific OIB.

A speech recovery experiment was performed to test our modified online technique. The setup for the experiment mainly consisted of a sound source (loudspeaker), a vibrating object (paper), and a high-speed camera as shown in Fig. 5. An input sound consists of fluctuations in air pressure at the surface of the vibrating object. These fluctuations cause the object to move, resulting in a pattern of surface displacement over time. In order to avoid contact vibrations, the loudspeaker was placed separate from the surface holding the object. Considering the limitation of the exposure time, a light-emitting diode (LED) photography lamp was used to provide

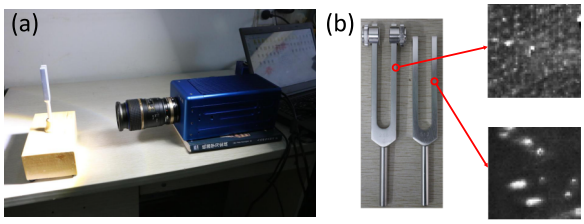

FIG. 2. Experimental setups (a) and sub-images (b) in tuning fork test.

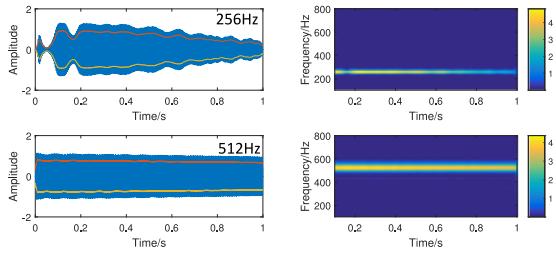

FIG. 3. Sound recovery results in tuning fork test.

enough illumination at approximately 0.5 m away from the object. Standard audio file reciting the phrase "University of science and technology of China" in Chinese was played by a loudspeaker at loud volume of about 80 dB (an actor's stage voice). The vibrating object was filmed at a distance of 1 m for about 3.35 s. Considering that the frequency range of male voice is generally from 50 Hz to 800 Hz, video frame rate was set to be 5 kHz with a resolution of 150 × 100 pixels during the experiment. Sub-images of 100 × 50 pixels with a handwritten "USTC" (abbreviation of the played speech) were selected in the filmed video. The video files were processed using Matlab R2015b on a machine with a single 3.40 GHz processor and 8 GB of RAM.

Fig. 6 gives the analysis of the original standard speech source and a series of speech recovery results under different conditions. We used the full length video, proceeding 1000 frames (the voice had not yet rang out) and proceeding 3000 frames (the voice had just rang out) of the video to obtain different OIBs and then recovered the speech information from the filmed video. In order to improve the distinguishability in spectrogram, Morlet wavelet instead of short-time Fourier transform (STFT) was used to analyse the time-frequency information of the speech signals.

It should be noted that the vibrations in speech recovery are subtle and much susceptible to environmental noise; the first component signal of the speech mainly consists of noise because of the higher energy of noise comparing with the fluctuations excited by the loudspeaker. Clear acoustic waves are all found in the second component signal (projections on the second OIB) of the video. Fig. 6(a) shows the source speech signal played by the loudspeaker and its spectrogram. The base frequencies and their high overtones are clearly shown in the spectrogram. Fig. 6(b) gives the sound recovery result by using the full length video, namely, the most time consuming recovery result. In Figs. 6(c) and 6(d), available

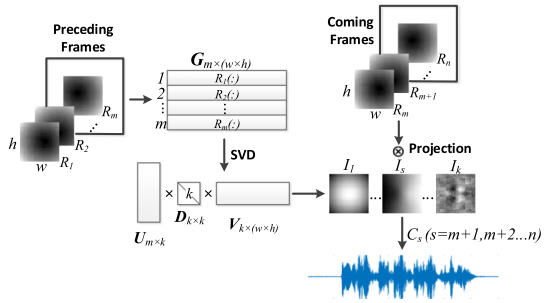

FIG. 4. Procedure of the online speech recovery.

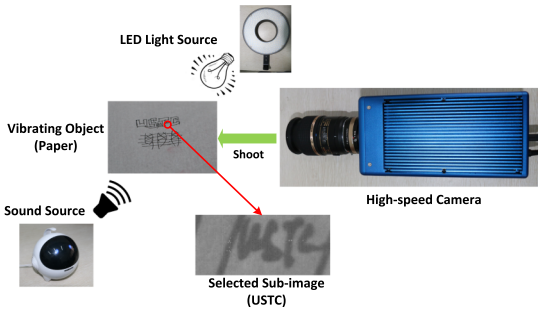

FIG. 5. Experimental setups in speech recovery experiment.

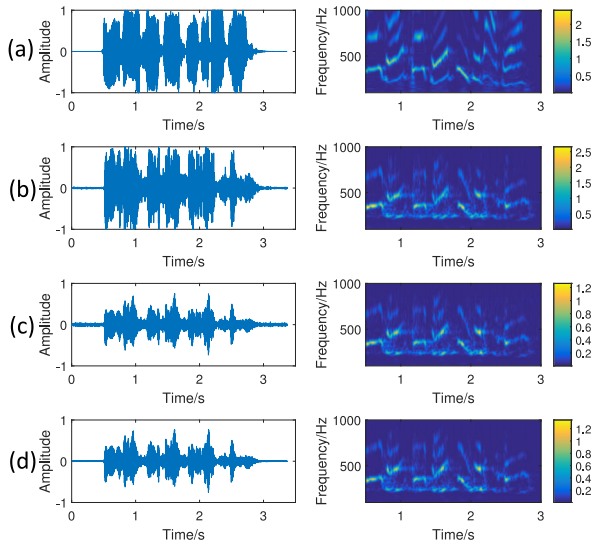

FIG. 6. Original source speech (a) and speech recovery results in different OIB conditions: (b) full length video; (c) 1000 frames; and (d) 3000 frames.

OIBs were calculated by using the preceding 1000 frames and 3000 frames, respectively. The rest of the frames were projected onto the specific second OIBs to compute speech signals online. We can see from the similar base frequencies and overtones in spectrograms that even using a small part of the video to build OIBs, the proposed online approach can still return understandable audio signals.

Quantitative analysis is given in Table I to evaluate the segmental signal-to-noise ratio (SSNR)[11] and intelligibility[12] of the results. The original speech signal is resampled to video rate at first for length matching. All the recovered speech signals are analyzed without any denoising or enhancement. It

TABLE I. Comparisons of the speech recovery results.

	SSNR (dB)	Intelligibility	Time (s)
Full length video	−4.71	0.37	43.17
1000 frames	−1.96	0.43	2.30
3000 frames	−2.11	0.41	7.14

is interesting to notice that the result that uses preceding 1000 frames of the video achieves the best SSNR and intelligibility. The reason of this phenomenon is that video projections on the OIBs calculated after the voice rang out reveal larger vibrations and also accompany with stronger noises. From the data of elapsed times, this 3.35 s speech is recovered from video within 1 min. Real-time sound measurement and display have been realized in the 1000 frames case. In general, as a sound recovery approach, the SVD-based method shows a great efficiency advantage. It should be noted that the environment noises and material property (such as density and rigidity) of the vibrating object can have a significant effect on the results. Signal denoising and enhancement will be further studied in our future work.

This research was supported by the Anhui Provincial Natural Science Foundation (Grant No. 1408085MKL83) and in part supported by the Key Technologies R&D Program of Anhui Province (Grant No. 1604a0902134).

[1]S. Rothberg, J. Baker, and N. A. Halliwell, J. Sound Vib. **135**, 516 (1989).
[2]P. K. Rastogi and P. Jacquot, Opt. Lett. **12**, 596 (1987).
[3]Z. Zalevsky, Y. Beiderman, I. Margalit, S. Gingold, M. Teicher, V. Mico, and J. Garcia, Opt. Exp. **17**, 21566 (2009).
[4]L. Zipser and H. Franke, in *Fifth International Conference on Vibration Measurements by Laser Techniques* (International Society for Optics and Photonics, 2002), pp. 192–198.
[5]M. Martarelli, P. Castellini, and E. P. Tomasini, Exp. Fluids **54**, 1 (2013).
[6]N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, ACM Trans. Graphics **32**, 80 (2013).
[7]A. Davis, M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand, and W. T. Freeman, ACM Trans. Graphics **33**, 79 (2014).
[8]J. G. Chen, N. Wadhwa, Y. J. Cha, F. Durand, W. T. Freeman, and O. Buyukozturk, J. Sound Vib. **345**, 58 (2015).
[9]A. Davis, K. L. Bouman, J. G. Chen, M. Rubinstein, F. Durand, and W. T. Freeman, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015), pp. 5335–5343.
[10]Z. Y. Wang, H. Nguyen, and J. Quisberth, Opt. Eng. **53**, 110502 (2014).
[11]J. H. Hansen and B. L. Pellom, in *International Conference on Spoken Language Processing* (CiteSeerX, 1998), Vol. 7, pp. 2819–2822.
[12]C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, IEEE Trans. Audio, Speech, Lang. Process. **19**, 2125 (2011).