Applied Acoustics 129 (2018) 316–321

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Double mode surveillance system based on remote audio/video signals acquisition

Tao Lv [a,b], He-yong Zhang [a,*], Chun-hui Yan [a,b]

[a] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, State Key Laboratory of Laser Interaction with Matter, Changchun 130033, China
[b] University of the Chinese Academy of Sciences, Beijing 10039, China

ARTICLE INFO

Article history:
Received 3 July 2017
Received in revised form 15 August 2017
Accepted 16 August 2017
Available online 20 August 2017

Keywords:
LDV
Laser hearing
Voice enhancement
Multimodal sensing

ABSTRACT

At present, remote human signature detection plays an increasingly important role in the field of anti-terrorism and security defense all around the world. In order to acquire remote, non-cooperative human signature signals, a double mode (audio/video) surveillance system is developed. The system mainly consists of a Laser Doppler Vibrometer (LDV), a pan-tilt-zoom (PTZ) camera and a theodolite. The LDV is used to acquire remote audio by detecting the vibration of the object (caused by the acoustic pressure). The PTZ camera is used to capture the video of the human body, and track the body when he/she moves, then analyze the image to select a proper vibrating target for LDV measurements. The theodolite is applied to control the orientation of the LDV. For the reason that many noise sources disturb the LDV-measured signals, such as laser speckle noises, environmental noises and the noises caused by sensor moving, a kind of speech enhancement technology (OM-LSA algorithm) is used to improve the intelligibility of the noisy voice signals detected by the LDV system. Experiments results indicated that the SNR and MOS of the LDV speech signal (the range is 150 m) can be increased by 135% and 26% respectively by using the speech enhancement technology, and the remote speech and video signals (the range is 30 m) can be obtained by the double-mode surveillance system.

1. Introduction

Remote human signature detection systems are widely deployed today for security purpose. In general, most remote human signature detection systems mainly depend on visual information [1]. Although video technologies (including visible and IR) have a great advancement in human signature detection, there is still a serious limitation in non-cooperative and hostile environments. The audio information, as an important data source, has not been fully explored yet. A few systems [2,3] have been reported to integrate visual and acoustic sensors, but in those systems, the acoustic sensors need to be close to the targets in monitoring. Parabolic microphones can be used for remote hearing and surveillance, which can capture voice at a fairly large distance in the direction pointed by the microphone. However, it is sensitive to the noise caused by wind and sensor motion. Laser Doppler Vibrometer (LDV) can measure extremely tiny vibration of a target at a long range [4–9]. On the other hand, objects near to the audio sources can be vibrated by the acoustic pressure. Therefore, the voice signals of a human could be acquired by capturing the vibration of a target's surface caused by the speech of the person next to the target. Li, Wang and et al [10–14] have presented their results in detecting and processing voice signals of people from large distances using a LDV from Polytec (includes a controller OFV-5000 with a digital velocity decode card VD-6, a sensor head OFV-505). However, in their work, the light of the LDV is 632 nm, which is not fit for realistic application due to its visible laser beam (can be perceived easily). Besides, the system is bulky and heavy because of the separated structure (the Polytec OFV 505 system has a size of 120 mm × 80 mm × 345 mm and weight of 3.4 kg).

In previous work, we have presented our results in detecting and processing voice signals of people from large distances using a self-designed LDV [15]. Now, a LDV with a configuration of all fiber is established. The all-fiber LDV system can offer the advantages of smaller size, lightweight design, robust structure, and the laser beam is invisible (1550 nm). Therefore, it is a better choice for remote speech detection. In this paper, the double-mode surveillance system based on self-design LDV, PTZ camera and theodolite is developed. The system can be used to detect remote human signature (visual and audio information). Several experiments are implemented to test its performance.

* Corresponding author.
E-mail address: zhanghy@ciomp.ac.cn (H.-y. Zhang).

http://dx.doi.org/10.1016/j.apacoust.2017.08.016
0003-682X/

2. Experimental setup

The principle block diagram of double-mode surveillance system is shown in Fig. 1. This system is composed of LDV, camera, theodolite and a personal computer.

The LDV is composed of transceiver (Fig. 2) and signal processor (Fig. 3). A 20-mW single-frequency (1550 nm) fiber laser with single longitudinal mode and narrow linewidth (less than 10 kHz) is used as the transmitter. The beam from the laser is divided into two beams by a 1×2 fiber coupler, one part is used for the local oscillator (LO) beam, and the other part acted as the transmitted beam. In order to discriminate the direction of target vibration, an acousto-optic frequency shifter (AOFS) is equipped to the LO beam. Afterwards, the LO is frequency-shifted up by 40 MHz using the AOFS, whose driving signal served as the reference signal input to the signal processing unit. The transmitted beam is focused on the target after passing through a circulator and a telescope. Due to the vibration of the target (vibration caused by the voice energy), the reflected beam carries Doppler frequency shift. This reflected beam is received by the same telescope, and it is mixed with LO by a 2×2 fiber coupler. Finally, the interference signal is detected by a balanced photodetector.

The output of the balanced photodetector is an FM signal with a center frequency f_{AOFS} of 40 MHz. In order to obtain acoustic signal, the demodulation methods are needed (Fig. 3). The detector output signal and reference signal are sampled by a dual-channel high-speed (250 M/s) data acquisition card. Then the detector output signal is divided into two equal parts, and the two divided signals mix with two orthogonal replicas of the reference signal, the corresponding in-phase (I) and the quadrature (Q) output signals can be obtained after through the low-pass filters. Finally, the arctangent phase function reconstructed the audio signals. Besides, the

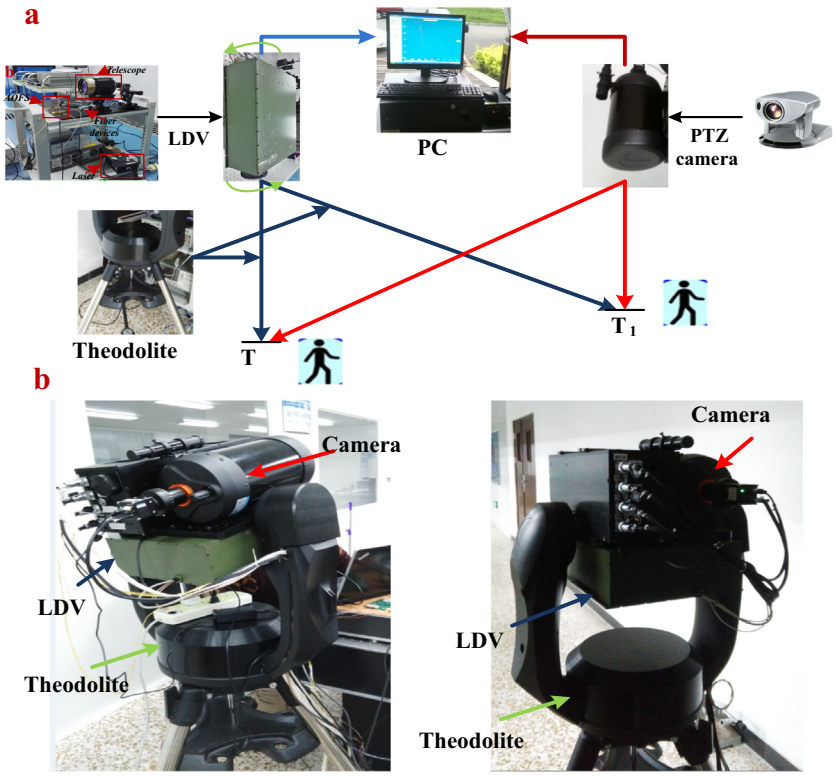

Fig. 1. (a) Schematic setup of the double-mode surveillance system. (b) The double-mode surveillance system.

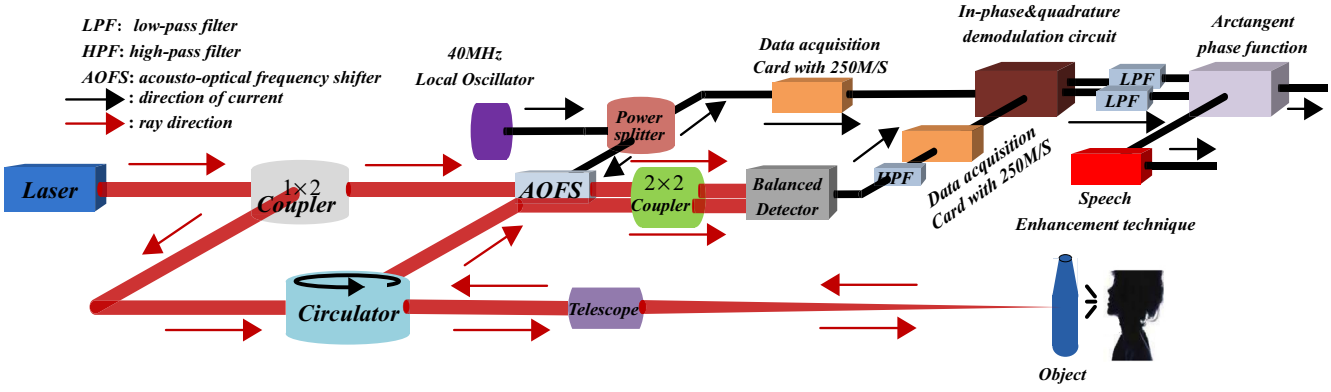

Fig. 2. Schematic diagram of the LDV transceiver.

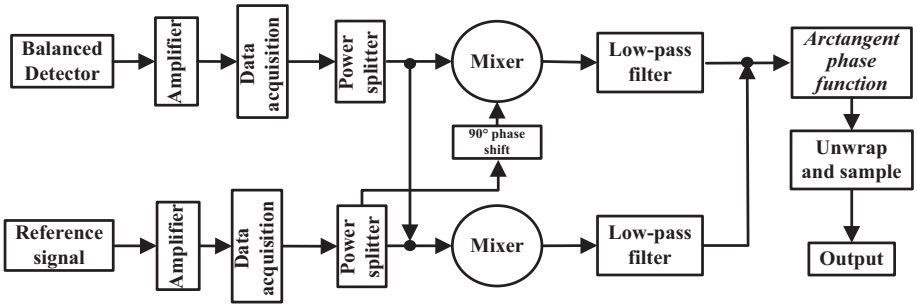

Fig. 3. Schematic diagram of the LDV signal processor.

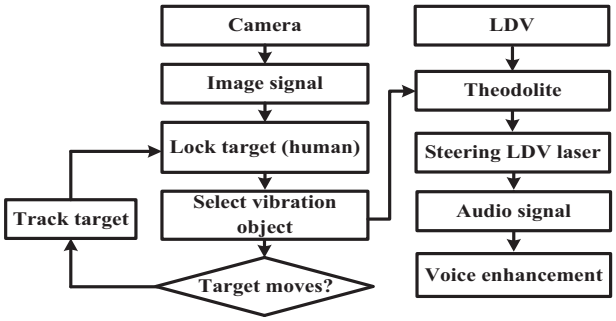

Fig. 4. Flowchart of the system.

ambiguity of the arctangent function can be removed by a phase unwrapping algorithm, which provides the integer number m, representing multiples of λ/4.

The PTZ camera has a 720 × 480 focal plane array and an auto-iris zoom lens that can change from 3.5 mm to 91 mm. The pan angle of the PTZ is −100° to 100°, and the tilt angle is −30° to 90°. The camera captures the image signal of the concerned target.

The theodolite system has a pan range from −130° to 130° and a tilt range from −60° to 70°. The rotation resolution is 12.36″.

The system implementation flowchart is shown in Fig. 4. The LDV which mounts on the theodolite acquires audio signals and PTZ camera collects image signals synchronously and send them to the PC to analyze the signals in real time. If the target moves, the PTZ camera tracks the target and aid the LDV to re-select a suitable vibration object (close to the target). In addition, it is important to note that the orientations of the LDV and PTZ camera are controlled separately, the orientation of the LDV is controlled by the theodolite, and the orientation of the PTZ camera is controlled by the PTZ.

3. Speech enhancement

The LDV system can detect remote acoustic signals effectively, but many noise sources disturb the LDV-measured signals, such as laser speckle noises, environmental noises, sensor motion and so on. The noise with frequency outside of normal speech frequency bandwidth can be filtered by a pass-band filter in a certain degree. However, the noise falling inside the voice frequency range still exists. Therefore, an OM-LSA algorithm [16] is used to further improve the intelligibility of the noisy voice signals.

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise, respectively, and let $y(n) = x(n) + d(n)$ be the LDV-measured signal. By using the short-time Fourier transforms (STFT) and the window function, we have $Y(l,k) = X(l,k) + D(l,k)$, where k and l respectively represent the frequency bin index and the frame index. Let the $H_0(l,k)$ and $H_1(l,k)$ indicate respectively speech absence and presence.

$$\begin{aligned} H_0(l,k) &= D(l,k) \\ H_1(l,k) &= X(l,k) + D(l,k) \end{aligned} \tag{1}$$

An estimator for the clean speech STFT signal $X(l,k)$ is traditionally obtained by applying a gain function to each time frequency bin, i.e., $\hat{X}(l,k) = G(l,k)Y(l,k)$. The OM-LSA estimator is

$$\begin{aligned} G(l,k) &= \{G_{H_1}(l,k)\}^{p(l,k)} \cdot G_{\min}^{1-p(l,k)} \\ G_{H_1}(l,k) &= \frac{\zeta(l,k)}{1+\zeta(l,k)} \exp\left(\frac{1}{2}\int_{v(l,k)}^{\infty} \frac{e^{-t}}{t} dt\right) \end{aligned} \tag{2}$$

where $G_{H1}(l,k)$ is a conditional gain function given $H_1(l,k)$, $G_{min} \ll 1$ is a constant attenuation factor, and $p(l,k)$ is the conditional speech presence probability. Denoting by $\zeta(l,k)$ and $\gamma(l,k)$ the a prior and a posteriori SNRs, so $v(l,k)$ can be written as

$$v(l,k) = \frac{\gamma(l,k)\zeta(l,k)}{1+\zeta(l,k)} \tag{3}$$

The prior SNRs $\zeta(l,k)$ can be estimated as

$$\hat{\zeta}(l,k) = \alpha G_{H_1}^2(l-1,k)\gamma(l-1,k) + (1-\alpha)\max\{\gamma(l,k) - 1, 0\} \tag{4}$$

The $S(l,k)$ represent the smoothed-version of the power spectrum of $|Y(l,k)|^2$, Let $S_{min}(l,k)$ denote the minimum value of $S(l,k)$ within a finite window of length D, and let $S_r(l,k) = S(l,k)/(B_{min}S_{min}(l,k))$, where B_{min} represents the noise-estimate bias. Then, the conditional speech presence probability $p(l,k)$ can be written as

$$I(l,k) = \begin{cases} 1, & \text{if } S_r(l,k) > \delta_1 \\ 0, & \text{if } S_r(l,k) < \delta_0 \\ \frac{\ln(S_r(l,k)) - \ln(\delta_0)}{\ln(\delta_1) - \ln(\delta_0)}, & \text{otherwise} \end{cases} \tag{5}$$

$$\hat{P}(l,k) = \alpha_p \hat{P}(l-1,k) + (1-\alpha_p)I(l,k) \tag{6}$$

where α_p is the smooth coefficient, δ_1 and δ_0 are represented as the upper threshold and lower threshold respectively.

In order to check the validity of the OM-LSA algorithm, the experiment is carried out (in the corridor close to lab) (see Fig. 5). In the experimental setup, a speaker is located at a distance of about 150 m from the LDV and 45 cm from the target (a mineral water bottle, without retro-reflective). Due to the speaker's voice, the mineral water bottle vibrates and the vibration frequency equals to that produced by the sound-field frequency. The voice signals can be acquired by detecting the vibration of the mineral water bottle. Based on the LDV measured speech signals, the OM-LSA algorithm is used to improve the quality of the voice signals.

To evaluate the performance of the speech enhancement by the proposed technique, both subjective and objective evaluations are implemented. The subjective evaluation is named Mean Opinion Score (MOS) evaluation criterion, and this evaluation standard is shown in Table 1. The objective evaluation includes two criteria,

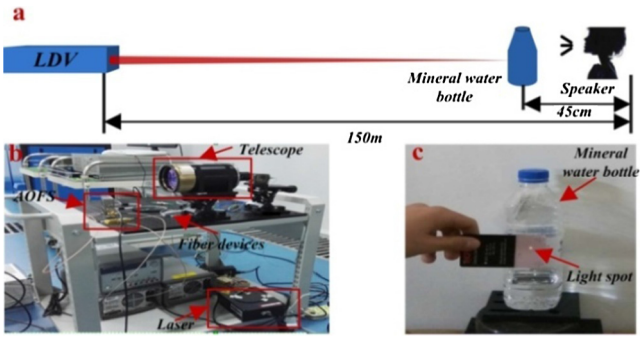

Fig. 5. (a) Experimental setup for detecting audio signals. (b) The all-fiber LDV system. (c) The target (a mineral water bottle without retro-reflective).

Table 1
The score standard of MOS.

The score of MOS	Intelligibility	Distortion level
5	Excellent	Don't feel distortion
4	Good	Just feel distortion
3	Okay	Feel distortion and have a little disgust
2	Difficult	Obviously feel distortion but Can bear it
1	Bad	Unacceptable

which are named spectrogram/waveform comparison and segmental signal–noise ratio (SNR) respectively.

Fig. 6 shows the spectrograms and waveforms of LDV speech signals(Fig. 6a), its enhanced signals (Fig. 6b), and corresponding clean signals(Fig. 6c) captured at the same time using a cell phone (the distance between the speaker and the cell phone is 20 cm).

It can be seen from the spectrograms and waveforms that the LDV speech signals (Fig. 6a) are contaminated by the noise seriously, the noise concentration distribution in the high frequency range (about 2000–3000 Hz), and relative weak noise in the low frequency part (under 500 Hz). Fig. 6b shows that the noise is largely attenuated by the OM-LSA algorithm, the spectrograms and waveforms are similar to the clean signals (Fig. 6c). This suggests that the OM-LSA algorithm can improve the quality of the noisy voice signals effectively. In addition, we use SNR values and MOS evaluation criterion (invite ten volunteers) to evaluate the performance of the speech enhancement by the proposed technique. The results are shown in Table 2. The MOS of the LDV speech signals is 2.7, and the SNR is 4.36 dB. The MOS of the enhanced signals is 3.4, and the SNR is 10.27 dB. The subjective and objective evaluations demonstrated the used speech enhancement technology can improve the quality of the noisy voice signals effectively.

Table 2
The SNR and MOS.

Method	SNR (dB)	MOS
LDV speech signals	4.36	2.7
The enhanced signals	10.27	3.4

4. Experimental results

In order to indicate the acquiring capability of remote speech signals and video signals, the experiment is carried out (in the lab). The lab was thought as a "non-cooperative" environment since all the objects are naturally placed there (Fig. 7a). When a human subject was detected in the scene (Fig. 7b), the system locked on the target and camera zoomed in to get a clearer image (Fig. 7c) and aid the LDV to select a suitable vibration object (close to the target). Finally, the system controlled the LDV to point the suitable vibration object (file box, Fig. 7d), and the audio signals were acquired and processed. When the target moved to another position, the system traced the target and obtained the image (Fig. 7e). Then, the captured image was analyzed to find a suitable vibration object. Finally, the LDV laser beam was redirected to the suitable vibration object (water bottle, Fig. 7f), and the audio signals were acquired and processed.

To evaluate the performance of the double-mode surveillance system (especially the acquired audio quality), the objective evaluations is implemented. The objective evaluation is named spectrogram/waveform comparison.

Fig. 8 shows the spectrograms and waveforms of LDV speech signals (the variation object is file box, Fig. 8a), and corresponding clean signals (Fig. 8b) captured at the same time using a cell phone. Fig. 9 shows the spectrograms and waveforms of LDV speech signals (the variation object is water bottle, Fig. 9a), and corresponding clean signals (Fig. 9b) captured at the same time using a cell phone.

It can be seen from the spectrograms and waveforms that the LDV speech signals (Figs. 8a and 9a) are close to the clean signals (Figs. 8b and 9b). Besides we use MOS evaluation criterion (invite ten volunteers) to evaluate the performance of the system. In

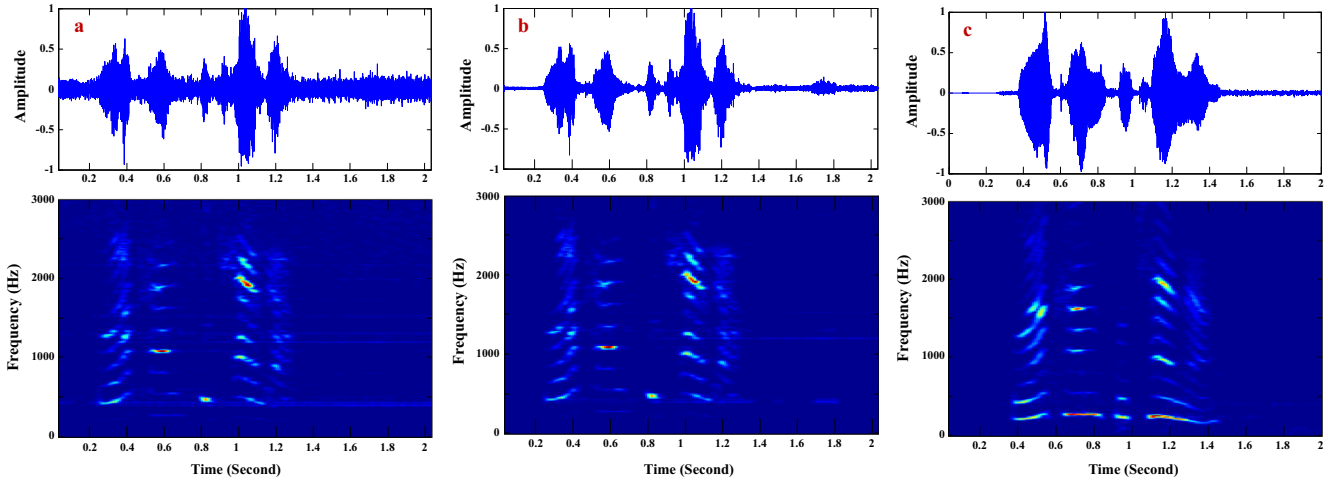

Fig. 6. Speech Spectrograms and waveforms. (a) Speech signals measured by the LDV. (b) Speech enhanced using the OM-LSA algorithm. (c) Clean speech signals measured by the cell phone.

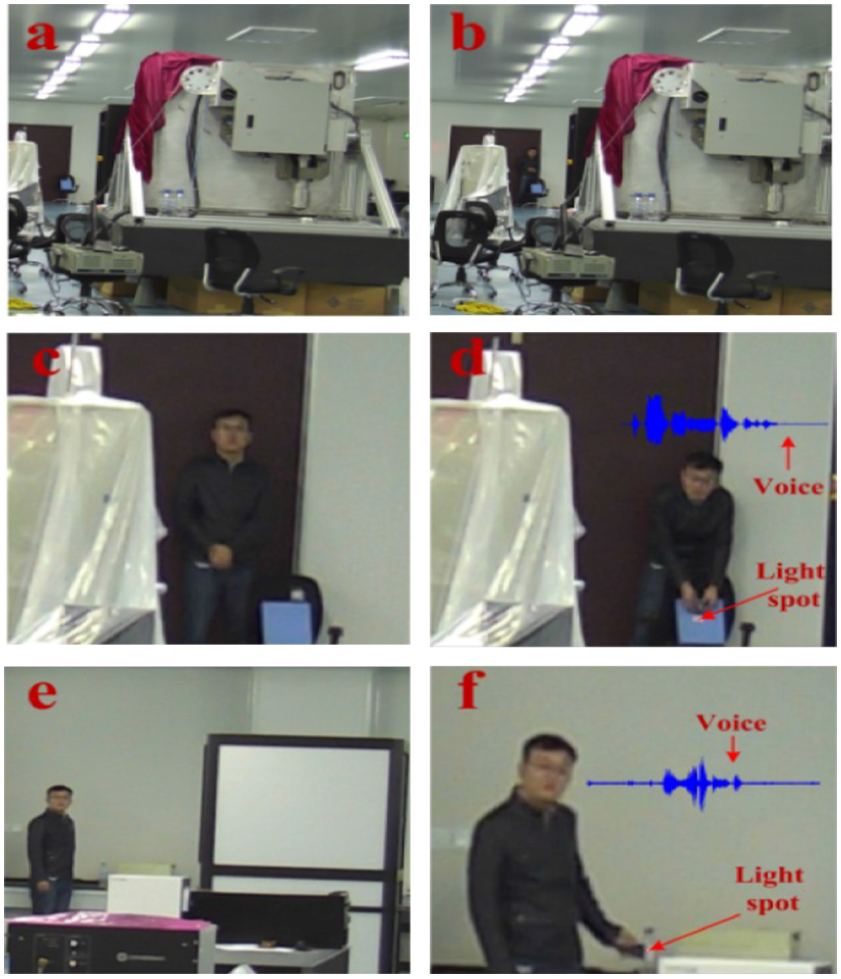

Fig. 7. Experiment results of remote A/V. (a) The lab image. (b) Finding a target in the lab. (c) Zoom in image. (d) Steer LDV to the suitable vibration object. (e) Target moves, the system track the target. (f) Steer LDV to the new suitable vibration object.

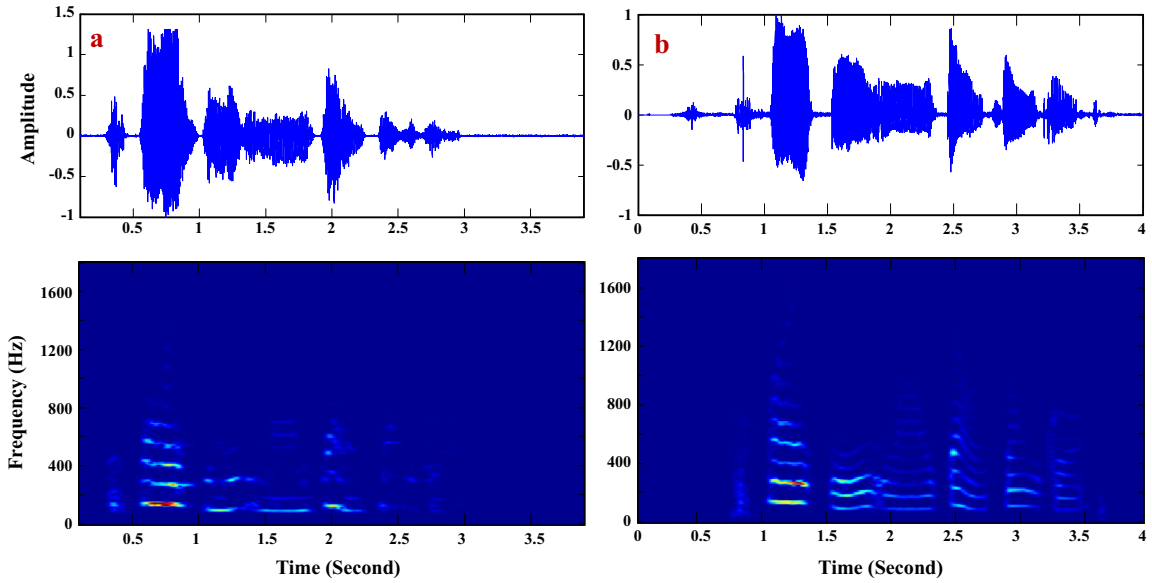

Fig. 8. Speech Spectrograms and waveforms. (a) Speech signals measured by the LDV (file box vibration). (b) Clean speech signals measured by the cell phone.

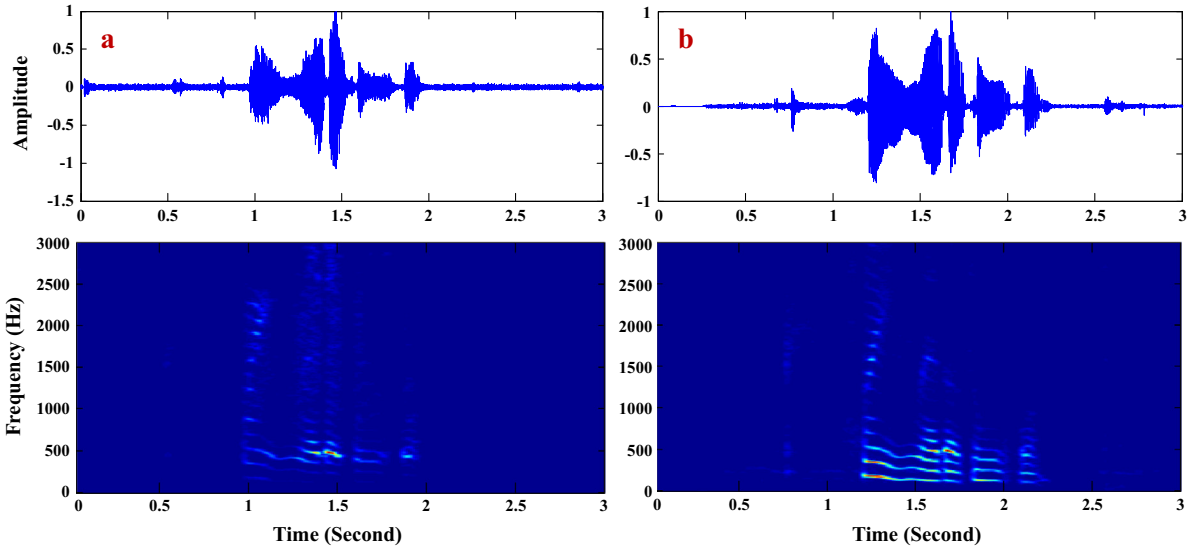

Fig. 9. Speech Spectrograms and waveforms. (a) Speech signals measured by the LDV (water bottle vibration). (b) Clean speech signals measured by the cell phone.

scenario 1 (Fig. 7d), the MOS of the LDV speech signals is 3.2, in scenario 2 (Fig. 7f), the MOS of the LDV speech signals is 3.0. Those results suggest that the system has ability to detect remote audio signals.

5. Conclusion

In conclusion, a double-mode surveillance system is developed to detect remote human signature (visual and audio information). This system has a self-design all-fiber LDV, a PTZ camera and a theodolite. The PTZ camera is used to remotely capture video signals for both visual information collection and the guidance of the LDV sensor to obtain the corresponding audio signals. The theodolite is used to control the orientations of the LDV. The LDV is used to acquire remote comprehensible speech signals. In addition, the audio signals enhancement technique (OM-LSA algorithm) is applied in this system to improve the quality of audio signals. Experiments' results indicated that the remote (30 m) speech signals and visual signals can be obtained by the double-mode surveillance system. This system can be used in various applications such as disaster rescue and remote area surveillance.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No. 61205143.

References

[1] Li X, Chen G, Ji Q, Erik B. A non-cooperative long-range biometric system for maritime surveillance. In: ICPR; 2008.
[2] Zotkin D, Dura swami R, Nanda H, Davis L. Multimodal tracking for smart videoconferencing. In: Second international conference on multimedia and expo, Tokyo, Japan; 2001.
[3] Zou X, Bhanu B. Tracking humans using multimodal fusion. In: The 2nd joint IEEE international workshop on object tracking and classification in and beyond the visible spectrum (OTCBVS'05), San Diego, CA, US, June 20; 2005.
[4] Rzasa John R, Cho Kyuman, Davis Christopher C. Long-range vibration detection system using heterodyne interferometry. Appl Opt 2015;54(20):6230–6.
[5] Chiu Mi-Hung, Chen W-Chou, Tan Ch-Tai. Small displacement measurements based on an angular-deviation amplifier and interferometric phase detection. Appl Opt 2015;54(10):2885–90.
[6] Li Rui, Madampoulos Nicholas, Zhu Zhigang, Xie Liangping. Performance comparison of an all-fiber-based laser Doppler vibrometer for remote acoustical signal detection using short and long coherence length lasers. Appl Opt 2012;51(21):5011–8.
[7] Zhang Xin, Diao Weifeng, Liu Yuan, Zhu Xiaopeng, Yang Yan, Liu Jiqiao, et al. Eye-safe single-frequency single-mode polarized all-fiber pulsed laser with peak power of 361 W. Appl Opt 2014;53(11):2465–9.
[8] Diao Weifeng, Zhang Xin, Liu Jiqiao, Zhu Xiaopeng, Liu Yuan, Bi Decang, et al. All fiber pulsed coherent lidar development for wind profiles measurements in boundary layers. Chin Opt Lett 2014;12(7):072801.
[9] Shang Jianhua, Zhao Shuguang, He Yan, Chen Weibiao, Jia Ning. Experimental study on minimum resolvable velocity for heterodyne laser Doppler vibrometry. Chin Opt Lett 2011;9(8):081201.
[10] Qu Y, Wang T, Zhu Z. Remote audio/video acquisition for human signature detection. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops; 2009, p. 66–71.
[11] Li W, Liu M, Zhu Z, et al. LDV Remote voice acquisition and enhancement. In: International conference on pattern recognition; 2006, p. 262–5.
[12] Wang AT, Zhu Z, Divakaran A. Long range audio and audio-visual event detection using a laser doppler vibrometer. Evol Bio-Inspired Comput: Theory Appl IV 2010;7704. 77040J–77040J-6.
[13] Qu Y, Wang T, Zhu Z. Remote audio/video acquisition for human signature detection. Cvpr'09 Biometrics 2009:66–71.
[14] Qu Y, Wang T, Zhu Z. An active multimodal sensing platform for remote voice detection. In: Advanced intelligent mechatronics (AIM), 2010 IEEE/ASME international conference on; 2010, p. 627–32.
[15] Zhang He-yong, Lv Tao, et al. The novel role of arctangent phase algorithm and voice enhancement techniques in laser hearing. Appl Acoust 2017;126:136–42.
[16] Cohen Israel. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Process Lett 2002;9(4):113–6.