Optics and Laser Technology 111 (2019) 575–584

Contents lists available at ScienceDirect

Optics and Laser Technology

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

Full length article

A remote human activity detection system based on partial-fiber LDV and PTZ camera

Xiyu Han[a,b], Tao Lv[a,b,*], Shisong Wu[a,b], Yuanyang Li[a,b], Bin He[a]

[a] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China

HIGHLIGHTS

- A double-mode surveillance system is developed to detect remote human activities.
- A new LDV structure: partial-fiber structure is used to detect remote speech.
- A speech enhancement technique is applied to improve the quality of the voice.
- A YOLO algorithm is used to discriminate human target and surroundings.

ARTICLE INFO

Keywords:
Laser Doppler Vibrometer
Remote acoustic detection
YOLO algorithm
Multimodal detecting
Voice enhancement

ABSTRACT

To address the challenges of non-cooperative and remote human activity detection, a multimodal remote audio/video acquisition system is developed. The system mainly consists of a Pan-Tilt-Zoom (PTZ) camera and a Laser Doppler Virbometer (LDV). The traditional all-fiber structure has residual carriers, which degrades the system performance badly. To solve the problem, a partial-fiber LDV is developed to obtain remote audio by detecting the vibration of the object (caused by the acoustic pressure around the target). Besides, to improve the quality of LDV audio signals, a speech enhancement algorithm (OM-LSA) is applied to remove noises in the LDV audio signals. The PTZ camera can provide remote visual information. We also use the YOLO algorithm to discriminate human from the photos which are updated from the PTZ camera continuously. That is the primary application of the YOLO algorithm. Moreover, the YOLO algorithm is used to recognize the objects around the target person by processing the video signals acquired by PTZ camera, which can aid the LDV in finding a suitable vibration target. In experiments, we show that the remote (50 m) speech signals and visual signals can be obtained by this surveillance system. That means this system has the ability to detect remote human activities.

1. Introduction

Event detection systems are widely deployed for security purposes currently. In general, almost all human activity detection systems work mainly at the visual level only [1], but other information modalities (such as audio) that can be easily obtained and used as complementary information remain underexplored. A few systems have been reported to integrate visual and acoustic sensors [2,3], but in these systems, the acoustic sensors need to be close to the targets under supervision. Furthermore, these types of sensors need to be fixed at pre-determined locations. If the targets move out of the detection range, they will not obtain any signals. Laser Doppler Vibrometer (LDV) can measure the extremely tiny vibration of a target at a long range [4–8]. And objects near to the audio sources can be vibrated by the acoustic pressure. Therefore, a human's voice signals can be acquired by capturing the vibration of the object's surface, which is caused by the speech of the person close to the target. Li, Wang and et al. [9–13] 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, the light of the LDV is 632 nm falling in a range of visible laser beam in their work. Because it can be perceived easily, it is not fit for practical application. And the sensor heads are bulky and heavy because of the inner separated structures of the commercial LDV systems (e.g., the Polytec OFV 505 system has dimension of 120 mm × 80 mm × 345 mm and weight of 3.4 kg). An all-fiber LDV system has advantages in smaller size, more lightweight design, and more robust structure, therefore it is less prone to structural vibrations

* Corresponding author at: Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China.
E-mail address: 18767120269@163.com (T. Lv).

https://doi.org/10.1016/j.optlastec.2018.10.035
Received 22 May 2018; Received in revised form 11 September 2018; Accepted 18 October 2018
0030-3992/

and more suitable for remote speech detection [14]. However, in previous studies [15,16], we found the traditional all-fiber structure has residual carriers which damage the system performance severely. Moreover, the signal quality acquired by the LDV is mainly determined by the reflection and the vibration properties of the selected object's surface nearby the target [17]. Unfortunately, it is hard for users to adjust the LDV manually to aim at a suitable vibration target with the laser beam. Those shortcomings obstruct the conventional LDV applied to the remote event detecting. To overcome the defects described above, we do some improvements. A partial-fiber LDV is developed to solve the residual carriers' problem. Besides, to solve the object selection problem, we integrate a partial-fiber LDV with a pan-tilt-zoom (PTZ) camera, which can offer visual information and aid the LDV in finding a suitable vibration object. Consequently, both video and audio signals are captured concurrently for remote human activity detection.

2. Remote speech collection and enhancement

2.1. Principle of the LDV

Because the LDV is a relatively new modality for the speech collection, we give a short introduction of LDV system in this section.

Compared with the traditional lens arrays LDV structure, the all-fiber structure has advantages in smaller size, more lightweight design, more robust structure, so the all-fiber LDV is more suitable to be chosen as a voice sensor for laser listening system.

2.1.1. Conventional all-fiber LDV

The schematic diagram of a traditional all-fiber LDV [18] is illustrated in Fig. 1. The LDV is composed of the optical unit and the electrical unit. In the LDV system, a coherent laser beam is divided into two beams by an optical fiber splitter, one part acted as the local oscillator (LO) beam, and the other part acted as the transmitted beam. In order to discriminate the vibration direction of the target, the LO beam was equipped with an acousto-optic frequency shifter (AOFS). The transmitted beam is focused on the target after passing through a telescope. Due to the target vibration caused by the voice energy, the reflect beam carries a Doppler frequency shift. By received through the telescope and coupled into the fiber circulator, the reflect beam is mixed with the LO beam in a polarization maintaining fiber coupler to produce a beat signal, which is converted into a voltage signal by a photoelectric balanced detector. The intermediate frequency (IF) signals will emerge when the voltage signal passes a band-pass filter. The LO signal, echo signal and IF signal are expressed as follows:

$$\begin{aligned} I_{LO} &= A_{LO}\cos[(\omega_c + \omega_{AO})t + \varphi_1] \\ I_S &= A_S\cos[\omega_c t + \varphi(t) + \varphi_2] \\ u_{IF} &= \alpha A_{LO}A_S\cos[\omega_{AO}t + \varphi(t) + \varphi_1 - \varphi_2] \end{aligned} \tag{1}$$

where A_{LO}, A_S are the amplitude of local-oscillator beam and signal beam respectively, ω_{AO} is the frequency shift caused by AOFS, φ_1 and φ_2 are random phases, α is the photoelectric conversion efficiency, $\varphi(t) = 4\pi S(t)/\lambda$ is the Doppler shift. $S(t)$ is the vibration displacement, λ is the wavelength.

The output of the balanced photodetector is an FM signal with a center frequency f_{AOFS}. In order to obtain acoustic signal, the demodulation methods are needed. Quadrature demodulation and arctangent phase algorithm is a classical method to demodulate the beat signal (as shown in Fig. 2) [19], This method has been analyzed and tested in recent years, which enable the realization of high-accuracy heterodyne signal processing in commercial LDV. The block diagram of the demodulation algorithm is depicted as Fig. 2. The IF signal is divided into two channels and mixed with two orthogonal (phase difference of 90°) replicas of the local oscillator. The corresponding in-phase (u_I) and quadrature (u_Q) demodulated signals are obtained after the low-pass filter, as shown in formula (2).

$$\begin{aligned} u_I &\approx \alpha A_{LO}A_S\cos[\varphi(t) + \varphi_1 - \varphi_2] \\ u_Q &\approx \alpha A_{LO}A_S\sin[\varphi(t) + \varphi_1 - \varphi_2] \end{aligned} \tag{2}$$

Moreover, based on the two orthogonal signals, the Doppler frequency shift $\varphi(t)$ can be calculated by using arctangent phase algorithm (as shown in formula (3)). Besides, the ambiguity of the arctangent function can be removed by phase unwrapping algorithm. Once the Doppler frequency shift $\varphi(t)$ is calculated, the speech signal can be reconstructed. It can be computed by using the following expression:

$$\varphi(t) = \arctan(u_Q/u_I) + m\pi + \Delta\varphi \tag{3}$$

2.1.2. Fiber circulator crosstalk

From the analysis for the optical structure of conventional all-fiber LDV, we can found that the fiber circulator is used to isolate the emergent signals from echo signals. However, lacking of the fiber circulator isolation will result in some emergent signals leaking in echo signals. Because of the existence of leakage of emergent signals, the echo and LO signals change their expressions:

$$\begin{aligned} I_{LO} &= A_{LO}\cos[(\omega_c + \omega_{AO})t + \varphi_1] \\ I_S &= A_S\cos[\omega_c t + \varphi(t) + \varphi_2] + A_E\cos[\omega_c t + \varphi_3] \end{aligned} \tag{4}$$

Therefore, baseband signals u_I and u_Q become the following expressions:

$$\begin{aligned} u_I &\approx \alpha A_{LO}A_S\cos[\varphi(t) + \varphi_1 - \varphi_2 - \varphi_3] + \alpha A_{LO}A_E\cos[\varphi_1 - \varphi_2 - \varphi_3] \\ u_Q &\approx \alpha A_{LO}A_S\sin[\varphi(t) + \varphi_1 - \varphi_2 - \varphi_3] + \alpha A_{LO}A_E\sin[\varphi_1 - \varphi_2 - \varphi_3] \end{aligned} \tag{5}$$

Assuming the phase difference *φ1-φ2-φ3* is equal to 0, based on formula 5, the Doppler frequency shift is obtained by adopting arctangent phase algorithm, as the following:

$$\begin{aligned} \varphi(t) &= \arctan\left(\frac{u_Q}{u_I}\right) = \arctan\left(\frac{\sin(\varphi(t))}{\cos(\varphi(t)) + \frac{A_E}{A_S}}\right) \\ &= \arctan\left(\tan(\varphi(t))\left(1 + (-1)^n\sum_{n=1}^{\infty}\left(\frac{A_E}{A_S}\right)^n\left(\frac{1}{\cos(\varphi(t))}\right)^n\right)\right) \end{aligned} \tag{6}$$

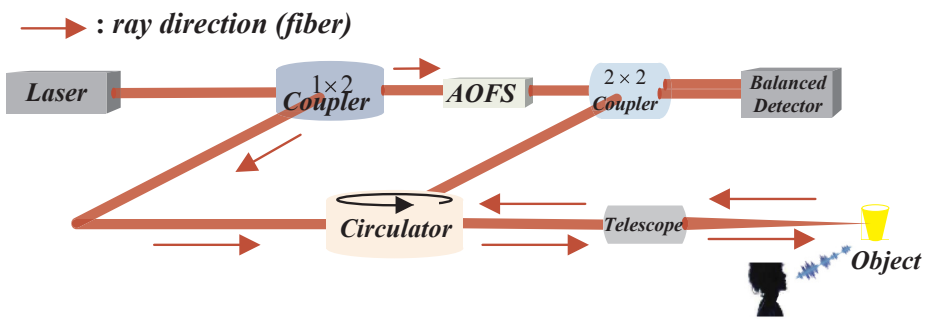

Fig. 1. Schematic diagram of the conventional all-fiber LDV.

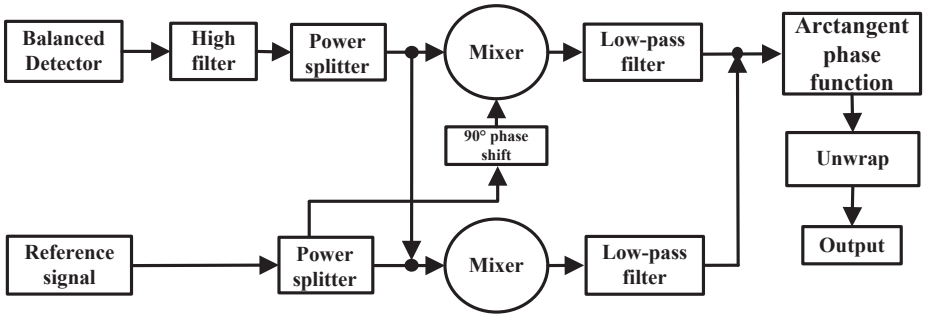

Fig. 2. Schematic diagram of the LDV signal processor.

The crosstalk of the fiber circulator is negligible when the intensity of the leakage of emergent signals are much less than echo signals $A_E << A_S$. However, when the two are of similar magnitude or the intensity of crosstalk is greater than the echo signals, the leakage of the fiber circulator affects the measurement accuracy of Doppler frequency shift directly. Unfortunately, in reality, the detection range is often more than dozens of meters. According to laser radar range equation, the power of echo signals is weaker than the leakage of emergent signals.

2.1.3. Partial-fiber LDV

In order to isolate the emergent signals from echo signals completely, a polarization prism is applied to substitute the fiber circulator. Fig. 3 shows the partial-fiber LDV which consists of a single-mode linear polarization laser source, two polarization-maintaining fiber couplers, two fiber collimators, a polarization prism, a telescope, a balanced detector, a quarter-wave plate, a half wave plate and several polarization-maintaining fibers. Because of the polarization property of the laser source, the emergent beams only transmit the polarization prism. Then transmitted beam is focused on the target after passing through a quarter-wave plate and a telescope in turn. The echo signals are collected by the telescope, and they pass through a quarter-wave plate again. Because the echo signals transmit the quarter-wave plate twice, the polarization direction has turned 90°, and the echo signals can be reflected by the polarization prism. In this way, the propagation paths of emergent signals and echo signals are separated completely.

2.1.4. Performance comparison between all-fiber and partial-fiber

To verify the performance differences between all-fiber system and partial-fiber system, an experiment is set up. As depicted in Fig. 4, a water bottle is regarded as the target. It is forced to vibrate under the effect of the loudspeaker box, which is driven by single-frequency sinusoidal tone generated by signal generator. We can control the single tone frequency by adjusting the frequency of the signal generator (in this experiment, the frequency is 500 Hz). The all-fiber LDV and partial-fiber LDV transmitted the laser beam perpendicularly to the surface of the bottle so as to gain the optimum reflected signal with maximum information.

The experimental results are shown in Figs. 5 and 6, we can obviously find that the all-fiber LDV has a strong residual carrier which caused by the leakage of emergent signals (see Fig. 6a). This residual carrier interferes the demodulation results significantly (see Fig. 6b and c). However, the partial-fiber LDV has the ability to eliminate the residual carrier effectively (see Fig. 5a), and the reconstructed signal waveform and frequency are uniform with the single tone are generated by us (see Fig. 5b and c).

2.2. LDV audio collection and enhancement

To indicate the acquiring capability of partial-fiber LDV for the remote speech, an experiment is set up. The experimental set up is similar to Section 2.1.4. The major difference is that the excitation source is no longer a single tone signal but a voice. Besides, to study the environmental adaptability of the system, some common items including plastic bag, mineral water bottle and computer screen are tested.

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 and sensor motion. The noise with the frequency outside the normal speech frequency bandwidth can be filtered by a pass-band filter to a certain degree. However, the noise falling inside the voice frequency range still exists. Therefore, an OM-LSA algorithm [20] 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 $y(n) = x(n) + d(n)$ 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 represent the frequency bin index and the frame index respectively. Let the $H_0(l,k)$ and $H_1(l,k)$ indicate speech absence and presence respectively.

$$\begin{aligned} H_0(l,k) &= D(l,k) \\ H_1(l,k) &= X(l,k) + D(l,k) \end{aligned} \tag{7}$$

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., $\hat{X}X(l,k) = G(l,k)Y(l,k)$. The OM-LSA estimator is

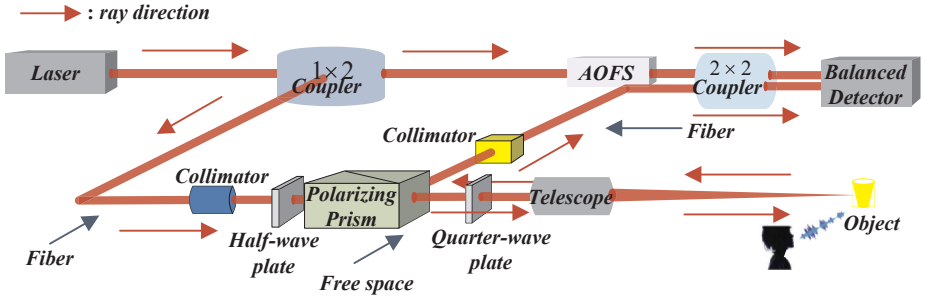

Fig. 3. Schematic diagram of the partial-fiber LDV.

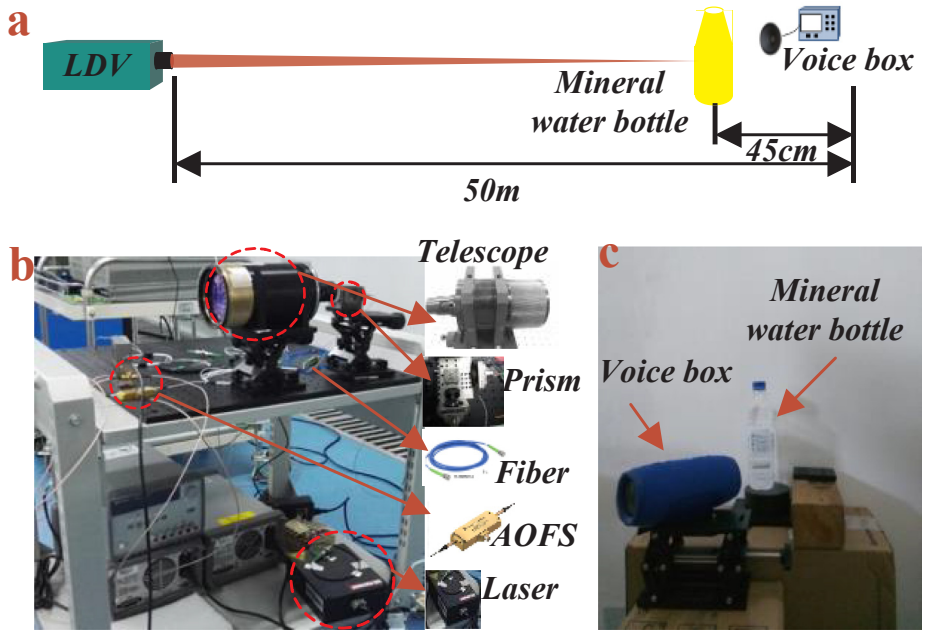

Fig. 4. (a) The experiment setup for single tone acquisition. (b) The LDV system. (c) The target.

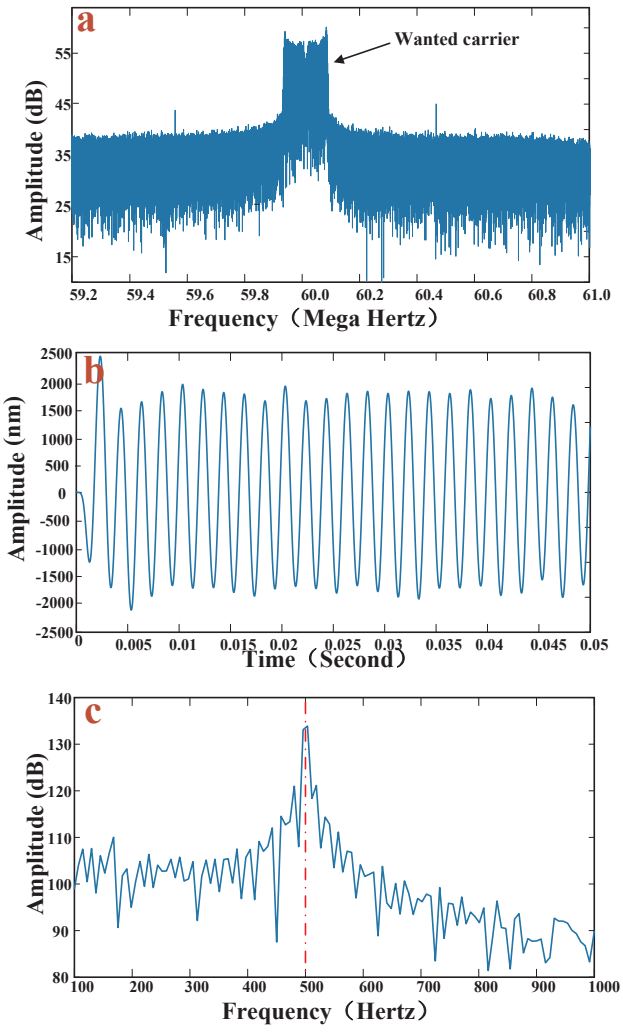

Fig. 5. Experimental results of the partial-fiber system.

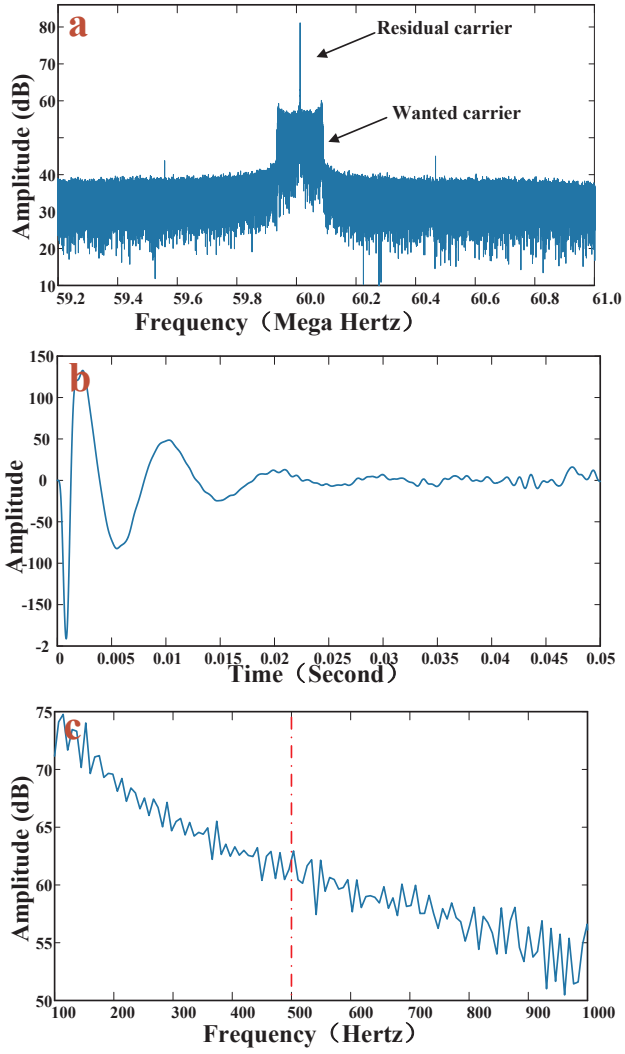

Fig. 6. Experimental results of the all-fiber system.

$$G(l,k) = \{G_{H_1}(l,k)\}^{p(l,k)} \cdot G_{\min}^{1-p(l,k)}$$
$$G_{H_1}(l,k) = \frac{\zeta(l,k)}{1+\zeta(l,k)} \exp\left(\frac{1}{2}\int_{\nu(l,k)}^{\infty} \frac{e^{-t}}{t} dt\right) \quad (8)$$

where GH1(l,k) is a conditional gain function given H1(l,k), Gmin≪1 is a constant attenuation factor, and p(l,k) is the conditional speech presence probability. Denoting by ζ(l,k) and γ(l,k) a prior and a posteriori SNRs, so ν(l,k) can be written as

$$\nu(l,k) = \frac{\gamma(l,k)\zeta(l,k)}{1+\zeta(l,k)} \quad (9)$$

where q(l,k) = P(H0(l,k)) is a priori probability for speech absence.

The prior SNRs ζ(l,k) can be estimated as

$$\hat{\zeta}(l,k) = \alpha G_{H_1}^2(l-1,k)\gamma(l-1,k) + (1-\alpha)\max\{\gamma(l,k)-1, 0\} \quad (10)$$

where S(l,k) represent the smoothed-version of the power spectrum of |Y(l,k)|2, Smin(l,k) denotes the minimum value of S(l,k) within a finite window of length D, and let Sr(l,k) = S(l,k)/(BminSmin(l,k)), where Bmin represents the noise-estimate bias. Then, the conditional speech presence probability p(l,k) can be written as

$$I(l,k) = \begin{cases} 1, & if\ S_r(l,k) > \delta_1 \\ 0, & if\ S_r(l,k) < \delta_0 \\ \frac{\log(S_r(l,k)) - \log(\delta_0)}{\log(\delta_1) - \log(\delta_0)}, & otherwise \end{cases} \quad (11)$$

$$\hat{P}(l,k) = \alpha_p \hat{P}(l-1,k) + (1-\alpha_p) I(l,k) \quad (12)$$

where αp is the smooth coefficient, δ1 and δ0 are represented as the upper threshold and lower threshold respectively.

2.3. Experimental results and discussion

The spectrograms and waveforms of LDV speech signals are collected from three different targets, and their enhanced signal, and the corresponding original clean speech are depicted in Fig. 7. For three different targets, all voice clips are similar to the corresponding original clean speech. However, the vibration of each surface has a different vibrating characteristic and frequency response. These differences exhibit a loss of meaningful high frequency contents in the spectrogram of LDV speech. In addition, all the voice clips captured from different target are contaminated by noise, which is distributed throughout a frequency range of acoustic signal. These noises may be caused by the inherent “speckle” problem on a normal “rough” surface, circuit noises and environmental noises. Fortunately, we find that the OM-LSA algorithm can eliminate noise effectively, but not completely. So, an OM-LSA algorithm can improve the quality of the noisy voice signals effectively. (Therefore, we use the OM-LSA algorithm to improve the quality of the noisy voice signals.) Meanwhile, we use an objective evaluation named PESQ [21] to evaluate the performance of the LDV and the speech enhancement by the proposed technique. This evaluation method has the highest correlation with subjective evaluation results. The evaluation method PESQ can be written as

$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \quad (13)$$

where D_{ind} is the average disturbance value, A_{ind} is the asymmetric disturbance value and a_0、a_1 and a_2 are coefficients.

The PESQ using LDV-captured speech collected from three different targets and their enhancement speech are reported in Table 1.

Experiment results indicated the intelligible speech signals can be acquired by the partial-fiber LDV, and the OM-LSA algorithm can eliminate noise effectively. Meanwhile, we find different objects have different vibration responses, so choosing a suitable target is critical for LDV voice acquisition.

3. Video detection and analysis

The most important purpose of video detection is to discriminate human from its background which is continuously updated from the PTZ camera. Besides, we found that the signal quality acquired from the LDV is mainly determined by the vibration properties of the selected objects nearby the target, so the selection of vibration target for LDV signals collection is also an important purpose for video detection. To achieve the above objectives, a YOLO algorithm is applied to recognize the human and objects nearby which can aid the LDV in finding a suitable vibration target via the PTZ camera video.

3.1. YOLO algorithm

The You Only Look Once (YOLO) is a state-of-the-art, real-time object detection system. The use of YOLO for detecting objects was first proposed by Redmon J in 2015 [22], and it is used here to aid the LDV in finding a suitable vibration target. YOLO is a new end-to-end detection algorithm. Although YOLO also belongs to CNN (Convolutional Neural Network), it obscures the differences among CG (Candidate Generation), FE (Feature Extraction) and CV (Candidate Verification) in the detecting process. This algorithm applies a single neural network to the entire image. This network divides the image into several regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities (Fig. 8). This model has some advantages over classifier-based systems. It looks at the whole image at test time so its predictions are informed by the global context in the image. It also makes predictions with a single network evaluation unlike systems like R-CNN, which require thousands for a single image. This makes it extremely fast, more than 1000× faster than R-CNN and 100× faster than Fast R-CNN. The standard YOLO can detect 45 pictures per second and the fast YOLO detection speed reaches 155 pictures per second [23,24].

3.2. Video detection experiment

In order to verify the ability of the YOLO to recognize human targets and objects nearby visually, an experiment is set up. The first step is to build the training dataset. The dataset in this paper is self-made dataset and the main source of dataset is pictures available on the internet. These images contain people and some common and easy-to-vibrate objects, such as glass cups, bags and so on, which will help to improve the accuracy and real-time performance of the system during the training phase. To increase the amount of data in the training set and improve the generalization ability of the model, this paper mainly uses left and right flip to enhance the data. Then, according to the self-made dataset, we used the model pre-trained on the ImageNet 1000-class competition dataset from YOLOv2 to train our detection networks. The pre-trained model could reduce the training time obviously. Next, we conduct several sets of tests to observe the performance. The system need fit for most daily scenes, so we took dozens of videos as a test set, which contain various people’s actions in different life scenes. Taking the angle factors into the consideration, those videos include images from different angles. In addition, we randomly search the life scenes videos and photos from the network for testing and observe its performance.

The precision rate, recall rate and processing time are shown in Table 2. Experiment results indicated the YOLO algorithm has ability to detect human and objects nearby in real time. Run time and recognition rate (especially the human) both meet the actual needs. Besides, the recognition rate of objects near people meets the requirements basically.

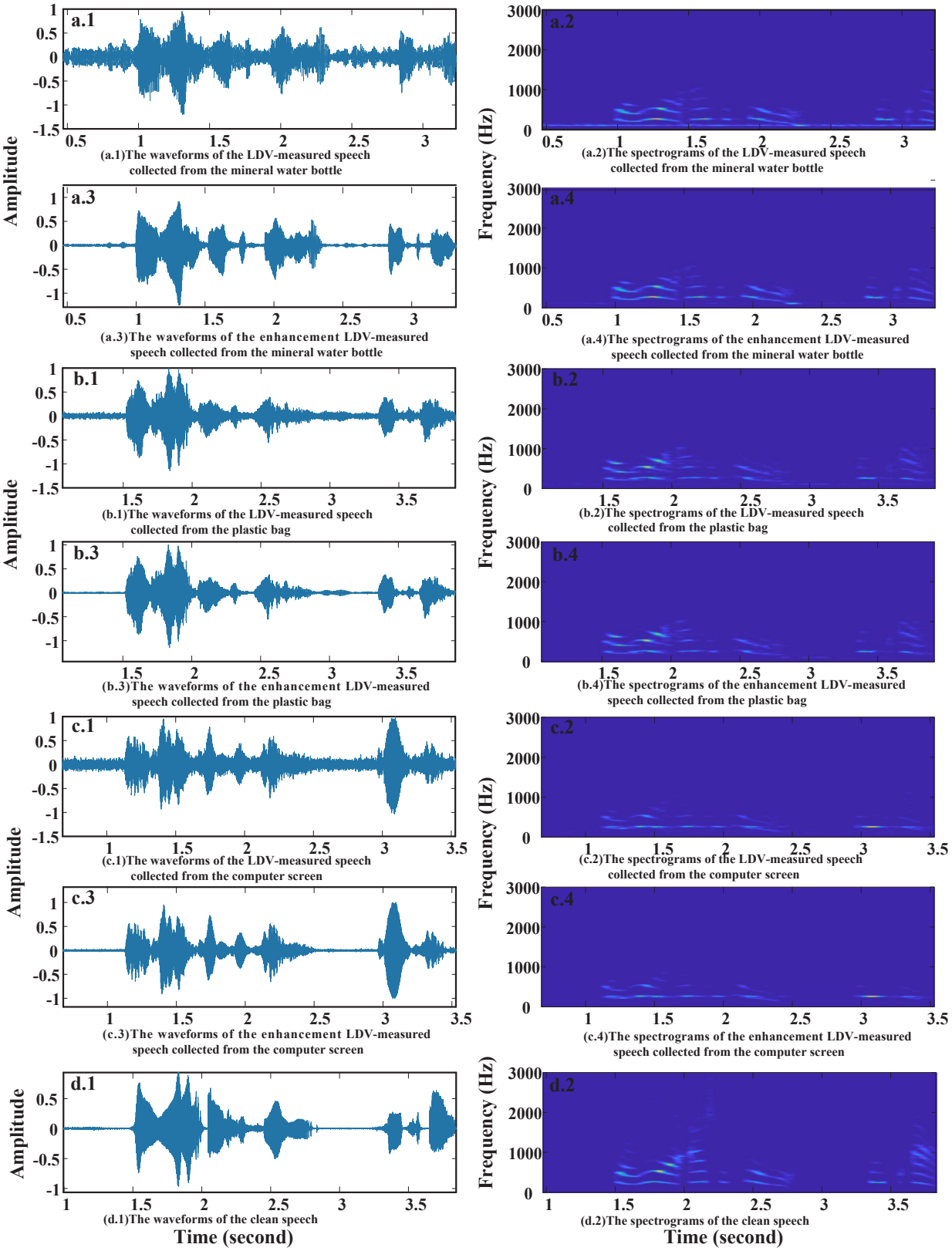

Fig. 7. Speech waveforms and spectrograms.

4. Integrated human activity detection system based on LDV and PTZ camera

Since the partial-fiber LDV has the ability to detect remote voice effectively, and the PTZ camera can capture video. Therefore we design a multimodal remote human activity detection system integrating the LDV together with PTZ camera. At the same time, we set up an experiment to verify the idea and achieved satisfactory results.

Table 1
The PESQ using LDV-measured speech collected from three different targets and enhancement speech.

Vibration target	LDV-captured speech	Enhancement speech
	PESQ(AVG)	*PESQ(AVG)*
Plastic bag	2.9213	3.5972
Mineral water bottle	2.1420	3.0831
Computer screen	1.8142	2.7029

4.1. System composition and experiment process

The principle block diagram and photos of the multimodal remote human activity detection system are shown in Fig. 9. This system is composed of partial-fiber LDV on a theodolite, a PTZ camera and a personal computer.

The use of the PTZ camera is to acquire visual information of targets at a large distance, obtaining the suitable image resolution of the targets with its zoom capability while keeping those targets inside the field of view (FOV) using its pan/tilt capability. The targets include both the human and surrounding objects here. The theodolite is used to control orientations of the LDV.

This 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. 10. The first step of system is to discriminate human targets from its background which is continuously updated from the PTZ camera by YOLO algorithm. When a possible human target is found, the PTZ first locks on to the human subject and zooms in to obtain a clear image of the subject. The ideal image should include both the human target and certain portion of its background. Meanwhile, the YOLO is used to recognize objects near human target, which can aid the LDV in finding a suitable vibration target. Then, the system controls the LDV laser beam to point onto the suitable vibration target in order to capture voice signals. Finally, the enhancement algorithm (in this paper, we use the OM-LSA algorithm) is used to improve the voice signals (If SNR of original voice signals is high enough, there is no need to use the enhancement algorithm of course). If the target moves, the PTZ camera will track the target and aid the LDV to re-select a suitable vibration object (using the YOLO algorithm to recognize the object). 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 its pan/tilt.

4.2. Experimental results and analysis

Some preliminary experimental results on remote video tracking and audio acquisition have been obtained in our lab (the distance is about 50 m). The lab was thought as a “non-cooperative” environment since all objects are placed there naturally (Fig. 11a). When a person

Table 2
The precision rate, recall rate and processing time.

	Human	Objects around people
Precision	99.89%	89.17%
Recall	96.49%	92.15%
Times(s)	0.0133	0.0133

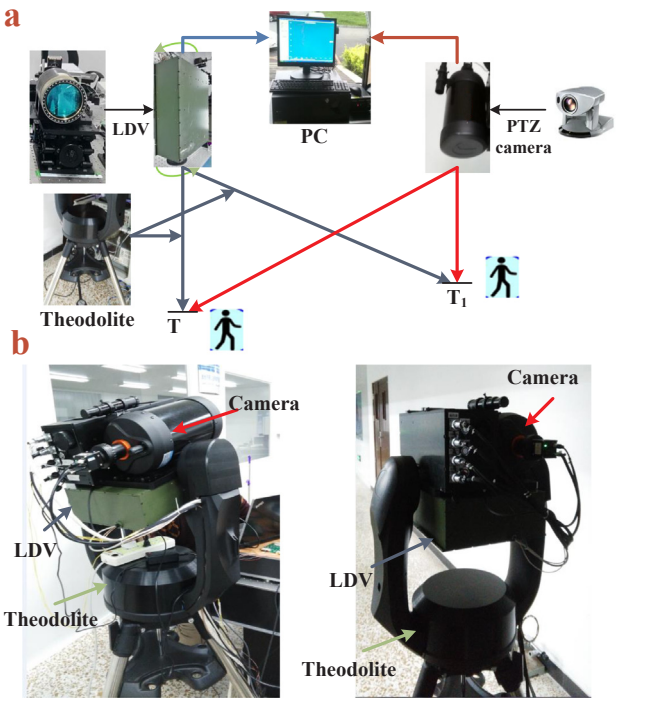

Fig. 9. (a) Schematic setup of the event detection system. (b) The prototype of the event detection system.

was detected in the scene, the system locked on and the camera zoomed in to get a clear image, with both the person and some surrounding objects (Fig. 11b). The objects in the image were recognized in order to find the best vibration target nearby the human (Fig. 11c). Finally, the system controlled the LDV to point at the suitable vibration object (Fig. 11d, paper bag), and audio signals were acquired and enhanced. When the human target moved to another position, the system traced the target and obtained the image (Fig. 11e). Then, the captured image was analyzed to find a suitable vibration object (Fig. 11f) again. Finally, the LDV laser beam was redirected to the suitable vibration object (Fig. 11g, TV screen), and the audio signals were acquired and enhanced again.

Throughout the experiment, the speed of the yolo recognition

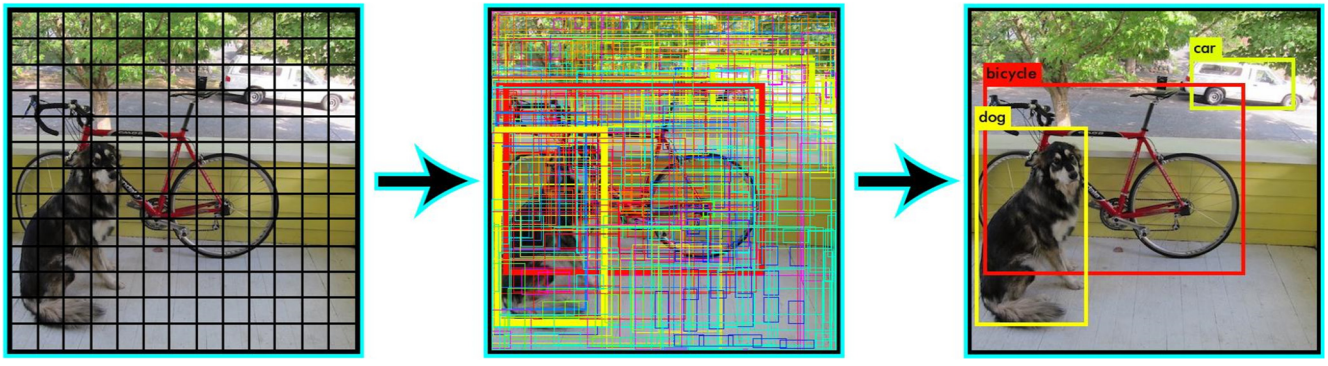

Fig. 8. The YOLO algorithm (this picture is from [24]).

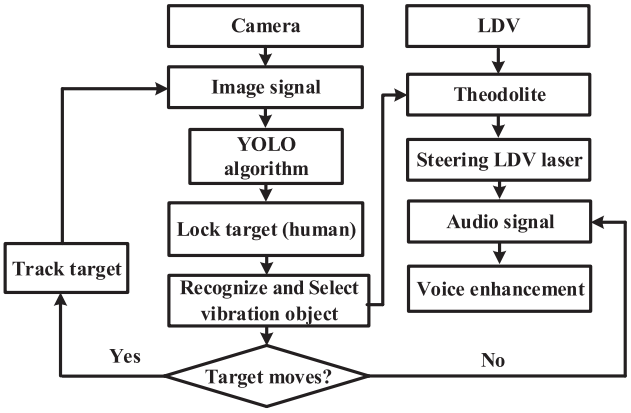

Fig. 10. The system implementation flowchart.

algorithm is 77 frames per second, and the target recognition rate of the human body is close to 100%. Besides, the objects around the human body can be well recognized (as shown in Fig. 11). The results show that the system can accurately identify targets and surrounding objects in real time.

To evaluate the quality of the audio captured by the double-mode surveillance system, the objective evaluations are implemented. The objective evaluation includes two criteria, which are named spectrogram/waveform comparison and *PESQ* respectively.

Fig. 12 shows the spectrograms and waveforms of LDV speech signals (the vibration object is a bag, Fig. 12a), and corresponding clean signals (Fig. 12b) captured by a cell phone at the same time. Fig. 13 shows the spectrograms and waveforms of LDV speech signals (the vibration object is a TV screen, Fig. 13a), and corresponding clean signals (Fig. 13b) captured by a cell phone at the same time.

It can be seen from the spectrograms and waveforms that the LDV

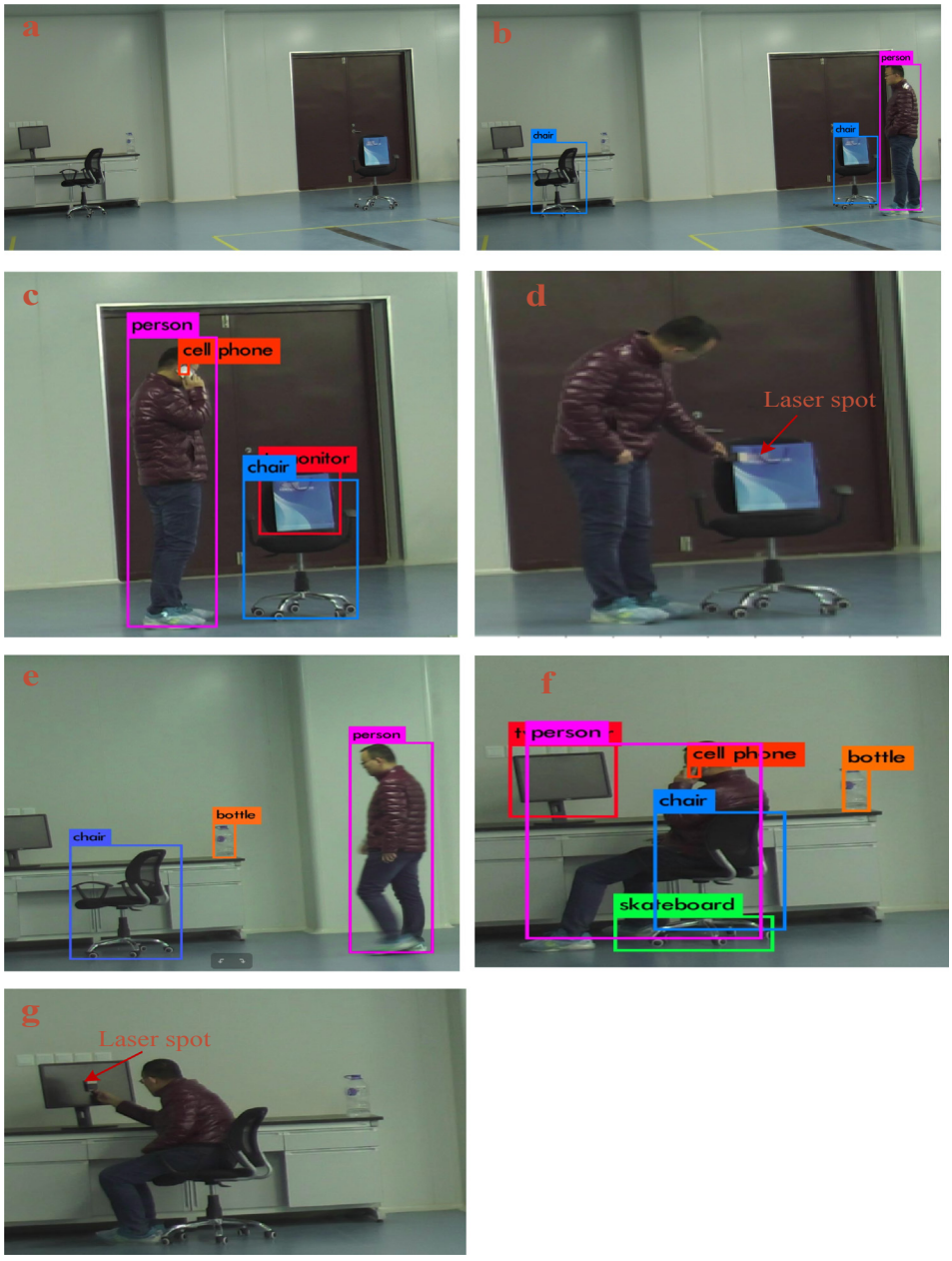

Fig. 11. Experiment results of remote A/V. (a) The lab image. (b) Finding a target in the lab. (c) Recognizing the objects and finding a suitable vibration target. (d) Steering LDV to the suitable vibration object (paper bag). (e) tracking the target, when the target moves. (f) Recognizing the objects and finding a suitable vibration target again. (g) Steering LDV to the new suitable vibration object (TV screen).

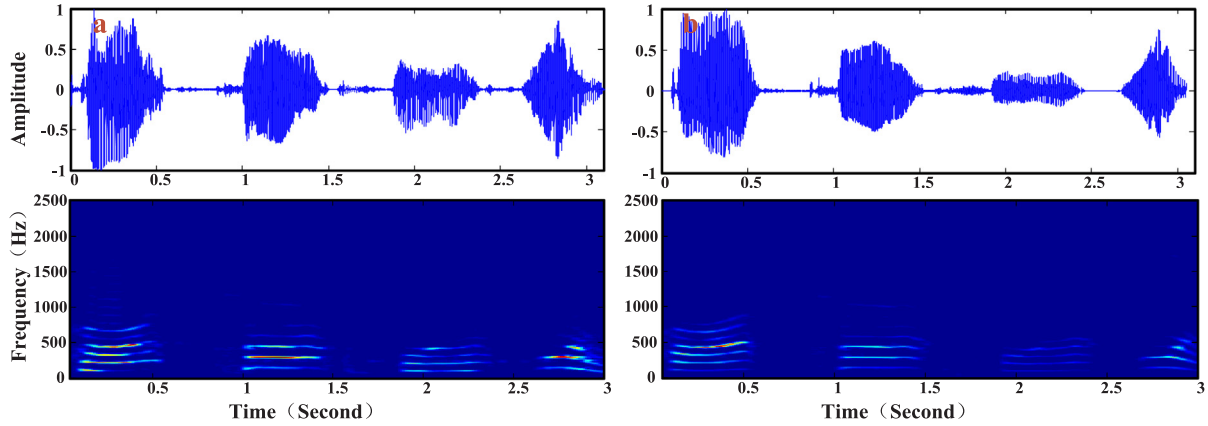

Fig. 12. Speech spectrograms and waveforms. (a) Speech signals measured by the LDV (bag vibration). (b) Clean speech signals measured by a cell phone.

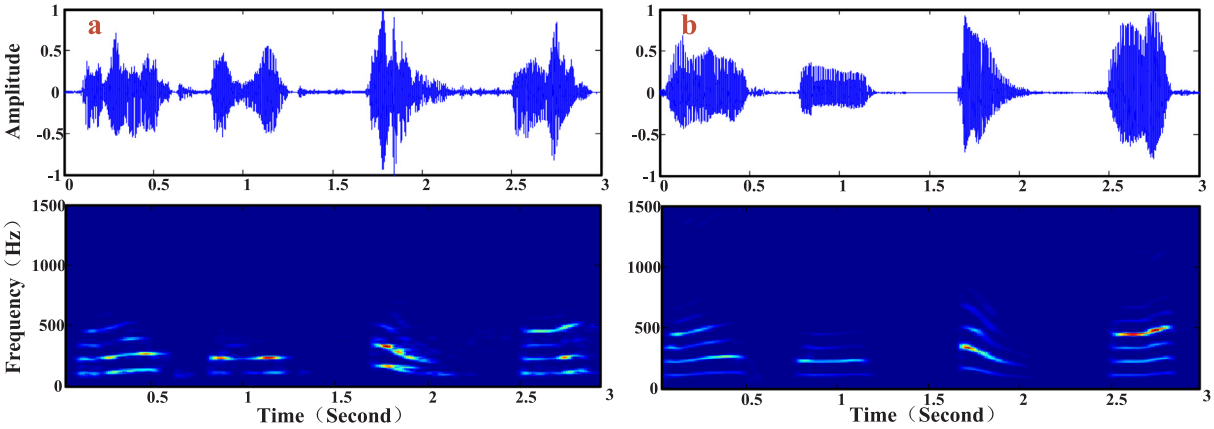

Fig. 13. Speech Spectrograms and waveforms. (a) Speech signals measured by the LDV (screen vibration). (b) Clean speech signals measured by a cell phone.

speech signals (Figs. 12a and 13a) are close to the clean signals (Figs. 12b and 13b). Besides, we used *PESQ* to evaluate the performance of the system. In scenario 1 (Fig. 11c), the *PESQ* of the LDV speech signals is 3.3516, and the *PESQ* of the LDV speech signals is 2.7429 in scenario 2 (Fig. 11f). Those results suggest that the system has the ability to detect remote audio signals.

In short, the entire experiment results prove that the system can detect remote audio and visual signals effectively.

5. Conclusion

In conclusion, a double-mode surveillance system is developed to detect remote human activities (combines visual information with audio information) using a partial-fiber LDV and a PTZ camera. The PTZ camera is used to capture video signals remotely. According to the visual information collected by PTZ camera, a YOLO algorithm is applied to discriminate human target. The partial-fiber LDV is applied to obtain the corresponding audio signals remotely by detecting the vibration of the object nearby audio sources. However, the quality of the voice acquired from the LDV is mainly determined by reflection and vibration properties of the selected vibration objects. Faced with this, the YOLO algorithm is also applied to recognize the objects around the target human to assist the LDV in finding a suitable vibration target. Furthermore, the detected speech signals may be corrupted by many noise sources, such as laser photon noises, target movements, and background acoustic noises (wind, engine sound, etc.). Therefore, the OM-LSA speech enhancement algorithm is used to remove noises in the LDV audio. Experiment results indicated that the intelligible speech signals and visual signals can be obtained by the double-mode surveillance system at a relative large distance (50 m). In this experiment, the mean *PESQ* score of the LDV measured audio is about 3.0473. On the basis of the PTZ video, the YOLO algorithm can identify the human target and the objects around the human quickly and correctly (the speed of the yolo recognition algorithm is 77 frames per second, and the target recognition rate of the human body is close to 100%). This system can be used in various applications such as disaster relief and remote area surveillance. In the future, we are also interested in human recognition using Face Recognition and Speaker Recognition technologies based on the collected data (video and speech).

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

References

[1] X. Li, G. Chen, Q. Ji, Erik. Blasch, A non-cooperative long-range biometric system for maritime surveillance, ICPR, 2008.
[2] D. Zotkin, R. Duraiswami, H. Nanda, L. Davis, Multimodal tracking for smart videoconferencing, Second International Conference on Multimedia and Expo, Tokyo, Japan, (2001).
[3] X. Zou, B. Bhanu, Tracking humans using multimodal fusion, The 2nd Joint IEEE International Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum (OTCBVS'05), San Diego, CA, US, (2005) June 20.
[4] John R. Rzasa, Kyuman CHO, Christopher C. DAVIS1, Long-range, vibration detection system using heterodyne interferometry, Appl. Opt 54 (20) (2015) 6230–6236.
[5] Ming-Hung Chiu, Wei-Chou Chen, Chen-Tai Tan, Small displacement measurements based on an angular-deviation amplifier and interferometric phase detection, Appl. Opt. 54 (10) (2015) 2885–2890.
[6] Xin Zhang, Weifeng Diao, Yuan Liu, Xiaopeng Zhu, Yan Yang, Jiqiao Liu, Xia Hou, Weibiao Chen, Eye-safe single-frequency single-mode polarized all-fiber pulsed laser with peak power of 361 W, Appl. Opt. 53 (11) (2014) 2465–2469.
[7] Weifeng Diao, Xin Zhang, Jiqiao Liu, Xiaopeng Zhu, Yuan Liu, Decang Bi, Weibiao Chen, All fiber pulsed coherent lidar development for wind profiles measurements in boundary layers, Chin. Opt. Lett. 12 (7) (2014) 072801.
[8] Jianhua Shang, Shuguang Zhao, Yan He, Weibiao Chen, Ning Jia, Experimental

study on minimum resolvable velocity for heterodyne laser Doppler vibrometry, Chin. Opt. Lett. 9 (8) (2011) 081201.
[9] 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, pp. 66–71.
[10] W. Li, M. Liu, Z. Zhu, et al. LDV remote voice acquisition and enhancement, in: International Conference on Pattern Recognition, 2006, pp. 262–265.
[11] A.T. Wang, Z. Zhu, Divakaran A. Long, Range audio and audio-visual event detection using a laser doppler vibrometer, Evolut. Bio-Inspired Comput.: Theory Appl. IV 7704 (2010) 77040J–77040J-6.
[12] Y. Qu, T. Wang, Z. Zhu, Remote audio/video acquisition for human signature detection, Cvpr'09 Biometrics (2009) 66–71.
[13] Qu Y, Wang T, Zhu Z. An active multimodal sensing platform for remote voice detection, in: 2010 IEEE/ASME International Conference on, Advanced Intelligent Mechatronics (AIM), 2010, pp. 627–632.
[14] Rui Li, Nicholas Madampoulos, Zhigang Zhu, Liangping Xie, Performance comparison of an all-fiber-based laser Doppler vibrometer for remote acoustical signal detection using short and long coherence length lasers, Appl. Opt. 51 (21) (2012) 5011–5018.
[15] Z.H.A.N.G. He-yong, L.V. Tao, et al., A The novel role of arctangent phase algorithm and voice enhancement techniques in laser hearing, Appl. Acoust. 126 (2017) 136–142.
[16] L.V. Tao, Z.H.A.N.G. He-yong, Y.A.N. Chun-hui, Double mode surveillance system based on remote audio/video signals acquisition, Appl. Acoust. 129 (2018) 316–321.
[17] Rui Li, Tao Wang, Zhigang Zhu, Wen Xiao, Vibration, characteristics of various surfaces using an LDV for long-range voice acquisition, IEEE Sensors J. 11 (6) (2011).
[18] Jianhua Shang, Shuguang Zhao, Yan He, Weibiao Chen, Ning Jia, Experimental study on minimum resolvable velocity for heterodyne laser Doppler vibrometry, Chin. Opt. Lett. 9 (8) (2011) 081201.
[19] M. Bauer, F. Ritter, G. Siegmund, High-precision laser vibrometers based on digital Doppler signal processing, Fifth International Conference on Vibration Measurements by Laser Techniques, 2002, pp. 50–61.
[20] Israel Cohen, Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator, IEEE Signal Process. Lett. 9 (4) (2002) 113–116.
[21] A.W. Rix, J.G. Beerends, M.P. Hollier, et al., Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, IEEE Proceedings 2 (2001) 749–752.
[22] J. Redmon, S. Divvala, R. Girshick, et al., You Only Look Once: Unified, Real-time Object Detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[23] J Redmon, A Farhadi, et al. YOLOv3: An Incremental Improvement, arXiv:2018:1804.02767.
[24] YOLO algorithm, < https://pjreddie.com/darknet/yolo/ >.