Acquirement and enhancement of remote speech signals*

LÜ Tao (吕韬)[1,2]**, **GUO Jin** (郭劲)[1], **ZHANG He-yong** (张合勇)[1], **YAN Chun-hui** (晏春回)[1,2], and **WANG Can-jin** (王灿进)[1]

1. Changchun Institute of Optics Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

2. University of Chinese Academy of Science, Beijing 100049, China

(Received 16 March 2017)

To address the challenges of non-cooperative and remote acoustic detection, an all-fiber laser Doppler vibrometer (LDV) is established. The all-fiber LDV system can offer the advantages of smaller size, lightweight design and robust structure, hence it is a better fit for remote speech detection. In order to improve the performance and the efficiency of LDV for long-range hearing, the speech enhancement technology based on optimally modified log-spectral amplitude (OM-LSA) algorithm is used. The experimental results show that the comprehensible speech signals within the range of 150 m can be obtained by the proposed LDV. The signal-to-noise ratio (*SNR*) and mean opinion score (*MOS*) of the LDV speech signal can be increased by 100% and 27%, respectively, by using the speech enhancement technology. This all-fiber LDV, which combines the speech enhancement technology, can meet the practical demand in engineering.

Document code: A **Article ID:** 1673-1905(2017)04-0275-4

DOI 10.1007/s11801-017-7059-9

Laser Doppler vibrometer (LDV) can measure extremely tiny vibration of a target at a long range[1-4]. On the other hand, objects in the vicinity of the audio sources can be vibrated by the acoustic pressure. These two aspects motivate our research in a new application of LDVs, namely remote voice detection from surrounding vibrated objects. Several approaches have been proposed to detect acoustic signals by using the LDV system. Wang[5] presented a novel multimodal remote audio and video acquisition system, which mainly consists of an LDV from Polytec including a controller OFV-5000 with a digital velocity decode card VD-6, a sensor head OFV-505 and a pan-tilt-zoom (PTZ) camera. This system can detect remote audio/video signals. However, the light of the LDV is visible. Using this system to acquire remote voice can be perceived easily. Besides, because of the separated structure, the system is bulky and heavy (the Polytec OFV 505 system has a size of 120 mm×80 mm×345 mm and weight of 3.4 kg). Avargel[6] presented a remote speech measurement system which utilized an auxiliary commercial LDV. However, the detection distance is only a few meters, and the light of the LDV is visible. Rui Li[7] developed an LDV using a short coherence length laser to detect remote acoustical signal.

In this paper, an LDV system using a single-frequency (1 550 nm) fiber laser with single longitudinal mode and narrow linewidth (less than 10 kHz) and other fiber components is developed to detect remote acoustic signals. Since the LDV-measured signals are disturbed by several noise sources, including laser speckle noises, environmental noises and sensor motion, the speech enhancement technique called as optimally modified log-spectral amplitude (OM-LSA) algorithm is applied to improve the intelligibility of the noisy voice signals.

The principle block diagram of the all-fiber LDV system is shown in Fig.1. This system is composed of transceiver unit and signal processing unit. A 20 mW single-frequency (1 550 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 acts as the local oscillator (LO) beam, and the other part acts 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 serves as the LO 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 (caused by the voice energy), the reflected beam carries Doppler frequency shift. This reflected beam is received by the same telescope, and it

* This work has been supported by the National Natural Science Foundation of China (No.61205143).

** E-mail: 18767120269@163.com

is mixed with LO beam by a 2×2 fiber coupler. Finally, the interference signal is detected by a balanced photodetector.

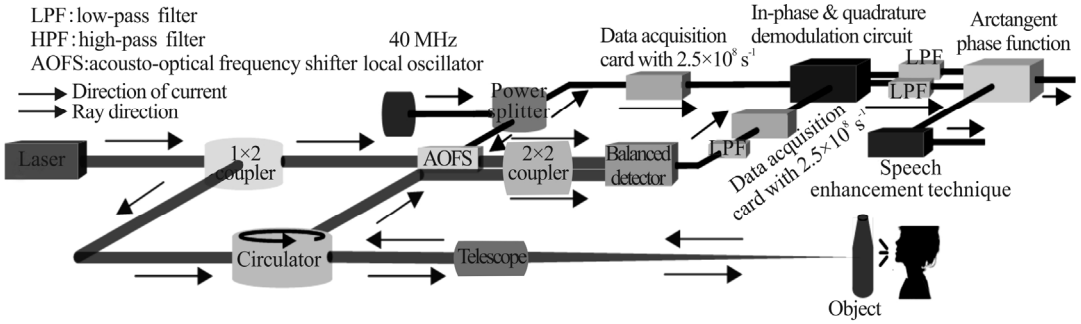

Fig.1 Schematic diagram of the LDV

The output of the balanced photodetector is a frequency modulation (FM) signal with a center frequency f_{AOFS} of 40 MHz. In order to get acoustic signal, the demodulation methods are needed, which are shown in Fig.2. The output signals from detector and LO are sampled by a dual-channel high-speed (2.5×10^8 s^{-1}) data acquisition card. Then the detector output signal is divided into two parts on average, and the divided signals mix with two orthogonal replicas of the LO signal, the corresponding in-phase (I) and the quadrature (Q) output signals can be obtained after passing the low-pass filters (LPFs). Finally, the arctangent phase function reconstructs the audio signals. Besides, the ambiguity of the arctangent function can be removed by a phase unwrapping algorithm, which provides the integer number m, representing the multiple of $\lambda/4$.

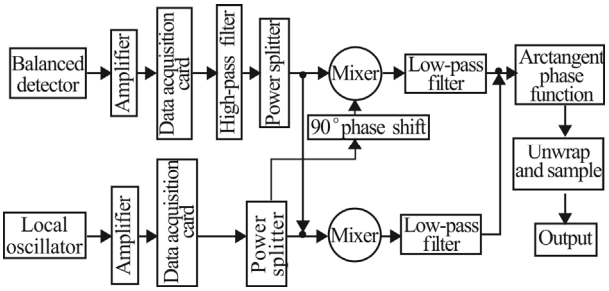

Fig.2 The block diagram of demodulation

The porposed 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 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[8] is used to further improve the intelligibility of the noisy voice signals.

Let $x(n)$ and $d(n)$ denote speech signal 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 $H_0(l,k)$ and $H_1(l,k)$ indicate respectively speech absence and presence, which are expressed as

$$H_0(l,k)=D(l,k), \tag{1}$$

$$H_1(l,k)=X(l,k)+D(l,k). \tag{2}$$

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

$$G(l,k)=\left\{G_{H_1}(l,k)\right\}^{p(l,k)}\cdot G_{\min}^{1-p(l,k)}, \tag{3}$$

$$G_{H_1}(l,k)=\frac{\zeta(l,k)}{1+\zeta(l,k)}\exp\left(\frac{1}{2}\int_{\nu(l,k)}^{\infty}\frac{\mathrm{e}^{-t}}{\mathrm{t}}\mathrm{d}t\right), \tag{4}$$

where $G_{H_1}(l,k)$ is a conditional gain function given $H_1(l,k)$, $G_{\min}<<1$ is a constant attenuation factor, and $p(l,k)$ is the conditional speech presence probability. Denoting the prior and posteriori signal-to-noise ratios (*SNR*) by $\zeta(l,k)$ and $\gamma(l,k)$, $\nu(l,k)$ can be written as

$$\nu(l,k)=\frac{\gamma(l,k)\zeta(l,k)}{1+\zeta(l,k)}. \tag{5}$$

The prior *SNR* $\zeta(l,k)$ and posteriori *SNR* $\gamma(l,k)$ can be estimated as

$$\zeta(l,k)=\alpha G_{H_1}^2(l-1,k)\gamma(l-1,k)+(1-\alpha)\max\left\{\gamma(l,k)-1,0\right\}, \tag{6}$$

$$\gamma(l,k)=\left|Y(l,k)\right|^2/\lambda_{\mathrm{d}}, \tag{7}$$

where $\lambda_{\mathrm{d}}=E\left[\left|D(l,k)\right|^2\right]$.

$S(l,k)$ represent the smoothed version of the power spectrum of $|Y(l,k)|^2$, $S_{\min}(l,k)$ denotes 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

$$p(l,k)=\alpha_p p(l-1,k)+(1-\alpha_p)I(l,k), \tag{8}$$

$$I(l,k)=\begin{cases}1, & \text{if } S_r(l,k)>\delta_1\\ 0, & \text{if } S_r(l,k)<\delta_0\\ \dfrac{\ln(S_r(l,k))-\ln(\delta_0)}{\ln(\delta_1)-\ln(\delta_0)}, & \text{otherwise}\end{cases}, \tag{9}$$

where α_p is the smooth coefficient, and δ_1 and δ_0 represent the upper threshold and lower threshold, respectively.

In order to indicate the capability of acquiring speech signals, the experiment is carried out in the corridor close to lab by using the all-fiber LDV system shown in Fig.3(b). In the experimental setup shown in Fig.3(a), a speaker is located with the distance of about 150 m from the LDV and 45 cm from the target (a mineral water bottle, without retro-reflection). As the speaker speaking, the mineral water bottle vibrates and the vibration frequency equals to that produced by the sound-field pressure. So the voice signals can be acquired by detecting the vibration of the mineral water bottle.

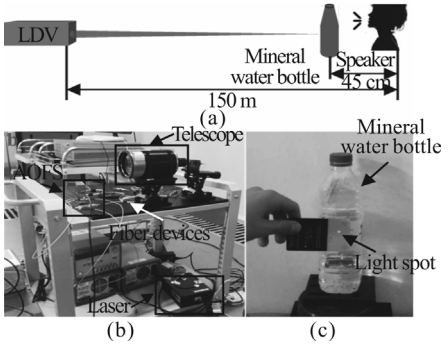

Fig.3 (a) Experimental setup for detecting audio signals; (b) The all-fiber LDV system; (c) The target (a mineral water bottle without retro-reflection)

To evaluate the performance of the LDV and the speech enhancement by the proposed technique, both subjective and objective evaluations are implemented. The subjective evaluation is named as mean opinion score (MOS) evaluation criterion, and this evaluation standard is shown in the Tab.1. The objective evaluation includes two criteria, which are named as spectrogram/waveform comparison and segmental *SNR*, respectively.

Tab.1 The evaluation score standard of *MOS*

Rating	Label
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Fig.4(c) shows the spectrograms and waveforms of LDV speech signals, and those of its enhanced signal and corresponding clean signals captured at the same time using a cell phone are shown in Fig.4(b) and (a), where the distance between the speaker and the cell phone is 20 cm.

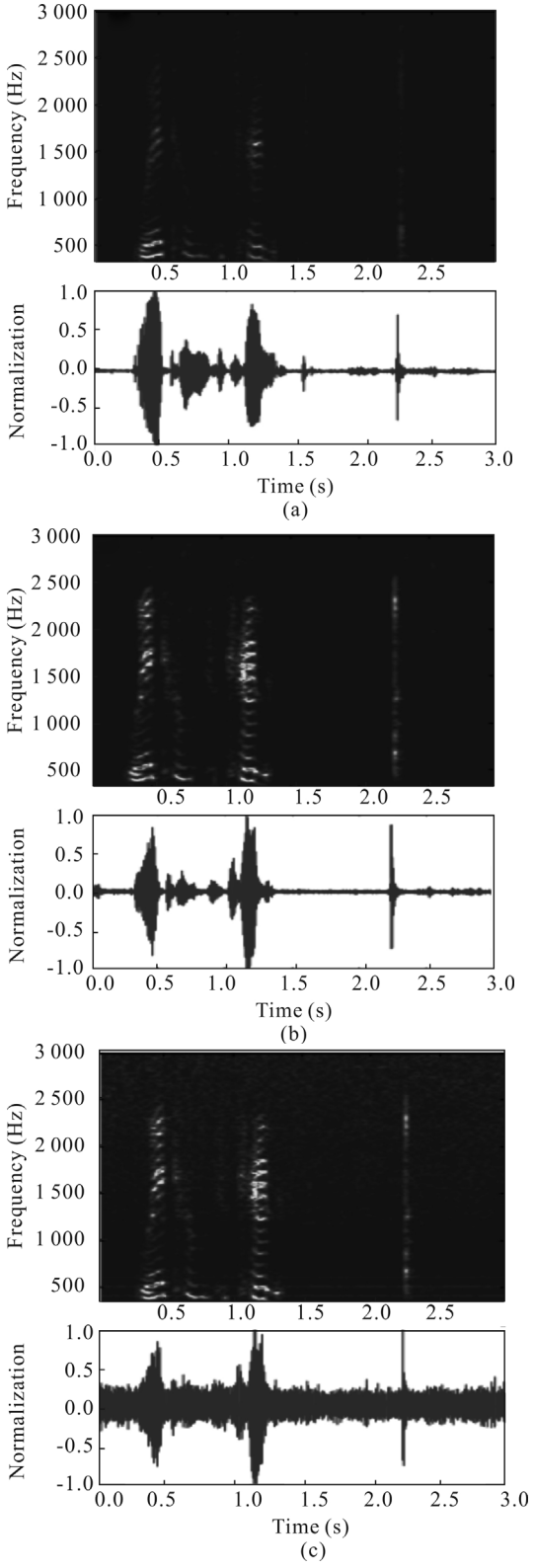

Fig.4 Speech spectrograms (upper) and waveforms (lower): (a) Clean speech signal measured by the cell phone; (b) Speech signal enhanced using the OM-LSA algorithm; (c) Speech signal measured by the LDV (All signals correspond to the speech of "Welcome to China".)

It can be seen from Fig.4(c) that the LDV speech signals are contaminated by the noise seriously. The high noise concentration distributes in the medium-high frequency range (about 1 500—3 000 Hz), and relative weak noise is in the low frequency part (under 500 Hz). However, the LDV speech signals are close to the clean signal shown in Fig.4(a), this suggests that the LDV has ability to detect the audio signals 150 m away. Fig.4(b) shows that the noise is largely attenuated by using the OM-LSA algorithm, and the spectrograms and waveforms are the closer to those of the clean signal. 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 (inviting ten volunteers) to evaluate the performance of the LDV and the speech enhancement by the proposed technique. The results are shown in Tab.2. The subjective and objective evaluation methods demonstrate the speech signals within the range of 150 m can be acquired by the LDV, and the used speech enhancement technology can improve the quality of the noisy voice signals effectively.

Tab.2 Results of *SNR* and *MOS* for two methods

Method	*SNR* (dB)	*MOS*
LDV speech signals	5.23	2.2
The enhanced signals	10.86	2.8

In conclusion, an all-fiber LDV system is developed to detect remote acoustic audio signals, the OM-LSA algorithm is applied in this system to improve the quality of signals. Experimental results indicate that the comprehensible speech signals within the range of 150 m can be obtained by LDV, and the OM-LSA algorithm can improve the intelligibility of the noisy voice signals detected by the LDV system effectively. In the domestic related research field[9-13], the detection distance of our LDV system is further. And this LDV will be a potential technology for the communication application in the future.

References

[1] LIU Li-sheng, ZHANG He-yong, WANG Ting-feng, GUO Jing and CHEN Chang-qing, Optics and Precision Engineering **23**, 1508 (2015). (in Chinese)

[2] J. R. Rzasa, K. Cho and C. C. Davis, Applied Optics **54**, 6230 (2015).

[3] Jianhua Shang, Shuguang Zhao, Yan He, Weibiao Chen and Ning Jia, Chinese Optics Letters **9**, 081201 (2011).

[4] Weifeng Diao, Xin Zhang, Jiqiao Liu, Xiaopeng Zhu, Yuan Liu, Decang Bi and Weibiao Chen, Chinese Optics Letters **12**, 080732 (2014).

[5] Qu Yufu, Wang Tao and Zhu Zhigang, An active Multimodal Sensing Platform for Remote Voice Detection, IEEE/ASME International Conference on Advanced Intelligent Mechatronics, 627 (2010).

[6] Avargel Y and Cohen I, Speech Measurements Using a Laser Doppler Vibrometer Sensor: Application to Speech Enhancement, Joint Workshop on Hands-free Speech Communication and Microphone Arrays, 109 (2011).

[7] Rui Li, Nicholas Madampoulos, Zhigang Zhu and Liangping Xie, Applied Optics **51**, 5011 (2012).

[8] Cohen I, IEEE Signal Processing Letters **9**, 113 (2002).

[9] Jianhua Shang, Yan He, Dan Liu, Huaguo Zang, and Weibiao Chen, Chinese Optics Letters **7**, 080732 (2009).

[10] Guo Bo, Qin Shui-jie and Tan Yi-dong, Journal of Optoelectronics·Laser **27**, 298 (2016). (in Chinese)

[11] Jian Zhou, Xiaoming Nie and Jun , Optics & Laser Technology **64**, 319 (2014).

[12] Li Hong-cai, Liu Chun-tong and Zhang Zhi-li, Journal of Optoelectronics·Laser **26**, 1902 (2015). (in Chinese)

[13] LV Tao, ZHANG He-yong, GUO Jing and YAN Chun-hui, Optics and Precision Engineering **25**, 569 (2017). (in Chinese)