# Research on separation and enhancement of speech micro-vibration from macro motion*

**CHEN Hong-kai** (陈鸿凯)[1,2], **WANG Ting-feng** (王挺峰)[1]**, **WU Shi-song** (吴世松)[1,2], **and LI Yuan-yang** (李远洋)[1]

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

*2. University of Chinese Academy of Sciences, Beijing 100059, China*

(Received 9 October 2019; Revised 27 November 2019)

Based on the 1 550 nm all-fiber pulsed laser Doppler vibrometer (LDV) system independently developed by our laboratory, empirical mode decomposition (EMD) and optimally modified Log-spectral amplitude estimator (OM-LSA) algorithms are associated to separate the speech micro-vibration from the target macro motion. This combined algorithm compensates for the weakness of the EMD algorithm in denoising and the inability of the OM-LSA algorithm on signal separation, achieving separation and simultaneous acquisition of the macro motion and speech micro-vibration of a target. The experimental results indicate that using this combined algorithm, the LDV system can functionally operate within 30 m and gain a 4.21 dB promotion in the signal-to-noise ratio (*SNR*) relative to a traditional OM-LSA algorithm.

**Document code:** A **Article ID:** 1673-1905(2020)06-0462-5

**DOI** https://doi.org/10.1007/s11801-020-9169-z

The relative movement along an optical axis between a target and light source can induce a Doppler frequency shift which represents the components of translation, vibration and rotation of the target. Since the laser Doppler vibrometer (LDV) system can measure these Doppler shifts in remote distance without contact and damage, the system has made great contributions in many research fields, such as target classification and recognition[1-3], general circulation and weather forecasting[4,5], mine detection[6,7], biometric feature research[8-10]. In addition, sound pressure can cause vibration of objects near the sound source. Therefore, the LDV system can be used for remote acoustic signal measurement[11-17]. Currently, several relevant researches based on LDV system are presented in order to obtain the micro acoustic vibration when there is no relative macro motion between target and light source. Avargel[11] presented a visible LDV system to detect the voice signal. However, this visible system has poor concealment and can be perceived easily. Lv acquired the acoustic vibration of a position-fixed object which is measured by an infrared LDV system[12] and raised an algorithm to eliminate speckles noise influence on system robustness[13]. Deng[15] and Peng[16] presented recognition algorithms to identify different speakers from LDV-captured speech, respectively. These researches on speech acquisition are based on a precondition that the position between object and LDV is fixed. However, there is still no effective solution for acoustic micro-vibration detection on targets with macro motion. In some specific applications, the macro motion and micro vibration of an object need to be acquired simultaneously, for example, to detect pedestrians talking on the road or people talking in moving vehicles. Generally, the vibration frequencies of human speech range from 300 Hz to 3 000 Hz, while the vibration amplitude ranges from nanometer to micrometer scale. By contrast, the macro motion of a target is a low-frequency signal, and the velocity of which is usually above the millimeter-scale per second. The mutual coupling between macro motion and speech micro-vibration dramatically extends the dynamic range of displacements so that the speech micro-vibration can be buried in the macro motion. Therefore, to obtain good comprehension of speech micro-vibration, it is essential and important to separate it from macro motion in the time domain. In this paper, the empirical mode decomposition (EMD) and optimally modified Log-spectral amplitude estimator (OM-LSA) algorithms are combined to obtain the macro motion and speech micro-vibration, simultaneously. The EMD algorithm is utilized to separate macro motion and the rough voice vibration of a target. Then, the high-quality enhanced speech is acquired by

* This work has been supported by the National Natural Science Foundation of China (No.61805234), the Key Research Program of Frontier Science, CAS (No.QYZDB-SSWSLH014), and the Foundation of State Key Laboratory of Laser Interaction with Matter (No.SKLLIM1704).

** E-mail: tfwang_ciomp@sina.com

utilizing the OM-LSA algorithm in rough voice vibration.

The whole system principle of the all-fiber pulsed LDV system is depicted in Fig.1. In the optical path shown in Fig.1(a), the continuous wave (CW) seed laser emits a single-mode, 1 550 nm laser beam, and the laser beam is split into a local oscillator (LO) beam and transmitted beam by a beam splitter (ratio 10/90). Ninety percent of the laser power conducts as the transmitted path with a 60 MHz frequency-offset modulated by acoustic-optic fiber shifter (AOFS). To generate pulsed light, the radio frequency driver of the AOFS is driven by applying a square wave pulse voltage with a 1 μm pulse width. Then, the pulsed transmitted beam passes through an optical circulator and a fiber collimator. The zoom telescope focuses the collimated pulse on a rough object surface, which is placed on an electronically controlled platform and vibrated by the sound from a loudspeaker. The platform moves along the optical axis. In the signal acquisition module, the echo of the transmitted beam mixing with a continuous LO path in the 3 dB fiber coupler is detected by a balanced detector and digitalized by an ADC.

In order to inverse and analyze the target-induced Doppler frequency shift, the echoes back reflected from the fiber end should be eliminated. The inversion process of demodulation is shown in Fig.1(b). The time-domain chopping technology is implemented to determine the echo of an object by calculating the peak of the echo of the fiber end and the time delay between the echo of the fiber end and the object. Then, the I & Q demodulation technique is utilized to shift the Doppler frequency containing 60 MHz carriers to the baseband. Two orthogonal signals, which are created by the numerically controlled oscillator (NCO), are multiplied by the time-domain chopping output, respectively. After passing through a low-pass filter, the error peak of the extraction in the non-pulse region for the time-domain chopping can be largely eliminated[17]. Finally, the arctangent algorithm and phase unwrapping algorithm are utilized to obtain the movement of the object which contains both macro motion and micro vibration.

The separation and enhancement process are shown in Fig.1(c). When there is a relative macro motion and micro-vibration between an object and the light source, the demodulated LDV signal of the object movement can be written as a summary of three components: $y(n)=x(n)+l(n)+d(n)$, where $x(n)$ represents the acoustic vibration characteristics of the object, $l(n)$ represents the relative macro motion and $d(n)$ represents the detected additive noise. This motion is usually a low-frequency signal compared with a speech acoustic signal, which generates a drifting trend for the speech vibration. If the traditional speech enhancement algorithm is directly utilized for an estimation, parts of $l(n)$ will be incorrectly identified as speech signal, and the macro motion component cannot be obtained. To separate and obtain the macro motion and eliminate the interference of the $l(n)$ component in the speech estimator, the EMD[18] algorithm is utilized in this paper. The EMD algorithm can adaptively and intuitively separate signals of different frequencies without any prior information from the original signal. By resolving the decomposed signal into a series of intrinsic mode functions (IMF) for different levels of fluctuations and a residual component, the EMD algorithm can simultaneously obtain the macro motion trend and micro vibration information. These IMFs can be considered as a modal space in which the high-frequency IMFs reflect the short-term fluctuation of the original signal, while the low-frequency IMFs and the residual component reflect the overall trend.

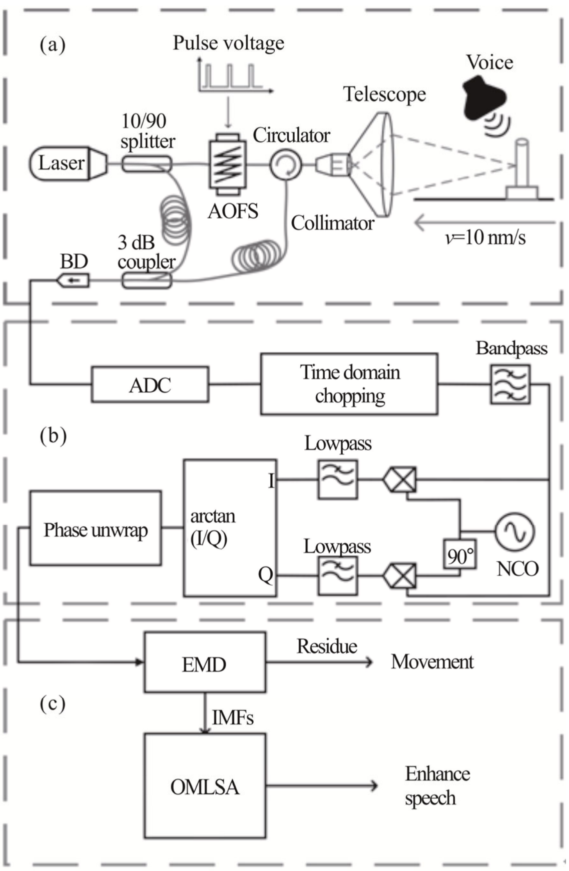

**Fig.1 Principle diagram of the all-fiber pulsed LDV system: (a) Optical layout; (b) Demodulation path; (c) Separation and enhancement path**

The extraction process of IMFs can be divided into two iterations of optimal IMF sifting and residual analysis. The process is described as follows.

Step 1: Let $y^{i,j}(n)$ denote the signal of the $j$th iterative sifting procedure for extracting the $i$th IMF, where $y^{1,1}(n)$ is the original data.

Step 2: Fit the lower and upper envelopes as $y_{\text{lower}}(n)$ and $y_{\text{upper}}(n)$ by interpolating, using a cubic spline, based on the local minima and maxima of $y^{i,j}(n)$.

Step 3: Calculate the mean curve $m^{i,j}(n)$ of two envelopes and obtain the refined evaluation:

$$h^{i,j}\left(n\right)=y^{i,j}\left(n\right)-m^{i,j}\left(n\right). \tag{1}$$

Let the input of the next sifting process be $y^{i,j+1}(n)$= $h^{i,j}(n)$ and repeat steps 1—3 until the evaluation coincides with one of the two IMF properties[18] that the evaluation mean equals zero or all the minima and all the maxima will be respectively negative and positive. Then, finish one iteration of optimal IMF sifting and obtain the $i$th finest IMF as:

$$IMF_i\left(n\right)=h^{i,j}\left(n\right). \tag{2}$$

Step 4: Subtract all the IMFs separated from the original data to obtain the next IMF extraction input by:

$$y^{i+1,1}\left(n\right)=y\left(n\right)-\sum_{k=1}^{i}IMF_k\left(n\right). \tag{3}$$

When $y^{i+1,1(n)}$ or $IMF_i$ is a monotonic curve or becomes less than the predetermined threshold, the iteration of the residual analysis is finished. The residue is defined as:

$$c\left(n\right)=y^{i+1,1}\left(n\right). \tag{4}$$

Therefore, original data can be rewritten by utilizing the decomposed $N$−1 IMFs and one residue as:

$$y\left(n\right)=\sum_{i=1}^{N-1}IMF_i\left(n\right)+c\left(n\right). \tag{5}$$

In the LDV system, the measurement noise, which involves speckle noise generated by the echo of a rough object, environmental random interference and electronic noise of the detection system, will seriously degrade the quality of the acquired speech signal. Thus, the enhancement of the speech signal serves as one of the critical operations of a speech detection system. The OM-LSA speech enhancement algorithm adopted in this paper, proposed by Cohen et al. in 2001[19], is an optimized algorithm for simultaneous estimation of noise and noisy speech signals and can maintain satisfactory performance in a low *SNR* environment.

Let $X(k,l)$ and $D(k,l)$ indicate the amplitude spectrum of the pure speech signal and uncorrelated additive noise, respectively. The amplitude spectrum of detected noisy speech can be written as $Y(k,l)$=$X(k,l)$+$D(k,l)$. Assume that there exists a specific spectral gain curve $G(k,l)$ obtained by speech enhancement filtering, and pure speech estimation can be obtained as follows:

$$\hat{X}\left(k,l\right)=G\left(k,l\right)Y\left(k,l\right), \tag{6}$$

where $l$ is the time frame index and $k$ is the frequency bin index. The OM-LSA algorithm gives two main assumptions for computing gain. Let $H_0(k,l)$=$D(k,l)$ and $H_1(k,l)$=$X(k,l)$+$D(k,l)$ represent the distribution of speech absence and presence, respectively, in $Y(k,l)$. A speech absence probability estimation is given to obtain the probability of frequency bin k existence $p(k,l)$ in the $l$th time frame index[17]. Then, the OM-LSA estimator is derived as follows:

$$G_{\mathrm{OM}}\left(k,l\right)=\left\{G_{\mathrm{H1}}\left(k,l\right)\right\}^{p(k,l)}G_{\mathrm{H0}}^{1-p(k,l)} \quad , \tag{7}$$

$$G_{\mathrm{H1}}\left(k,l\right)=\frac{\xi\left(k,l\right)}{1+\xi\left(k,l\right)}\exp\left(\frac{1}{2}\int_{\frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)}}^{\infty}\frac{\mathrm{e}^{-t}}{t}\mathrm{d}t\right), \tag{8}$$

$$G_{\mathrm{H0}}\left(k,l\right)<<1, \tag{9}$$

where $G_{\mathrm{H1}}(k, l)$ is a conditional gain function, $G_{\mathrm{H0}}(k,l)$ is the noisy attenuation constant factor and $\xi(k,l)$ and $\gamma(k,l)$ denote the priori and the posteriori *SNRs*, respectively.

To evaluate the performance of the proposed signal processing algorithm, three groups of target movement data are measured for comparison. The first group of data only contains uniform rectilinear movement at 10 mm/s. The other two groups of data have the same macro motion with the 500 Hz single-frequency harmonic vibration and speech of "Chang Chun" in Chinese, with the other experimental conditions invariant. The single-frequency harmonic vibration and speech are generated by the same speaker, and the distance between the target platform and the optical platform is 30 m. The measurement environment is shown in Fig.2.

The IMFs, which are decomposed from the demodulated signal by the EMD algorithm, are reunited to reconstruct the micro-vibration. The residue of the decomposed signal $c(n)$ represents the displacement of macro motion of the target. The decomposed components of the demodulated signal are illustrated in Fig.3. As shown in Fig.3, the demodulated signal is seriously distorted when there is micro-vibration. However, the residue still reveals macro motion. The reconstructed signal of Fig.3(a) shows the background noise. The reconstructed vibration in Fig.3(b) shows the single-frequency harmonic vibration of 500 Hz. Due to speckle and background noise, the vibration amplitude will float by approximately 20 μm. The reconstructed vibration in Fig.3(c) shows the rough acoustic signals reconstructed by the EMD algorithm which is also disturbed by noise.

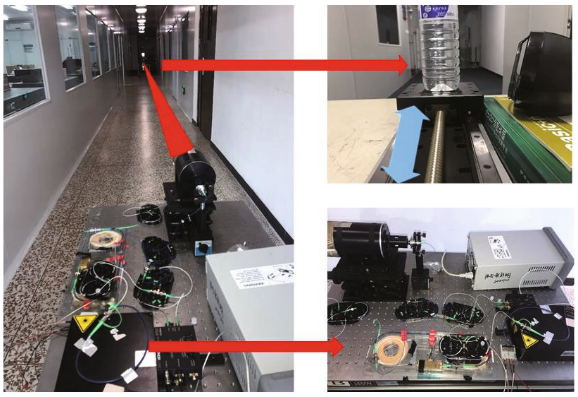

**Fig.2 The measurement environment and experimental setup of the LDV system**

To acquire the high-quality enhanced signal and verify the performance of proposed signal processing algorithm, OM-LSA algorithm is utilized on reconstructed signal reunited by EMD and on original demodulated LDV signal. The waveforms and spectrograms of clean speech are shown in Fig.4(a), and those of enhanced speech using the combined algorithm and the OM-LSA algorithm are shown in Fig.4(b) and Fig.4(c), respectively. Fig.4(b) shows that after enhancement of the reconstructed signal, the noise is largely attenuated, and the spectrograms are closer to those of the clean signal. When the OM-LSA algorithm is directly applied on the demodulated signal without reconstruction, which is shown in Fig.4(c), the macro motion trend will be incorrectly estimated as part of noise or speech and can not be separated from the micro speech vibration. The part of motion trend which is error estimated as speech will create error estimation of spectrogram, especially in the low frequency part. The spectrogram of enhanced signal by OM-LSA on demodulated signal depicted in Fig.4(c) has a poor performance at low frequencies (under 500 Hz) where more noise remains, compared with the enhanced signal by OM-LSA on reconstructed signal depicted in Fig.4(b). In addition, the *SNR*, MOS score of ten volunteers and average velocity of uniform rectilinear movement with speech vibration are shown in Tab.1.

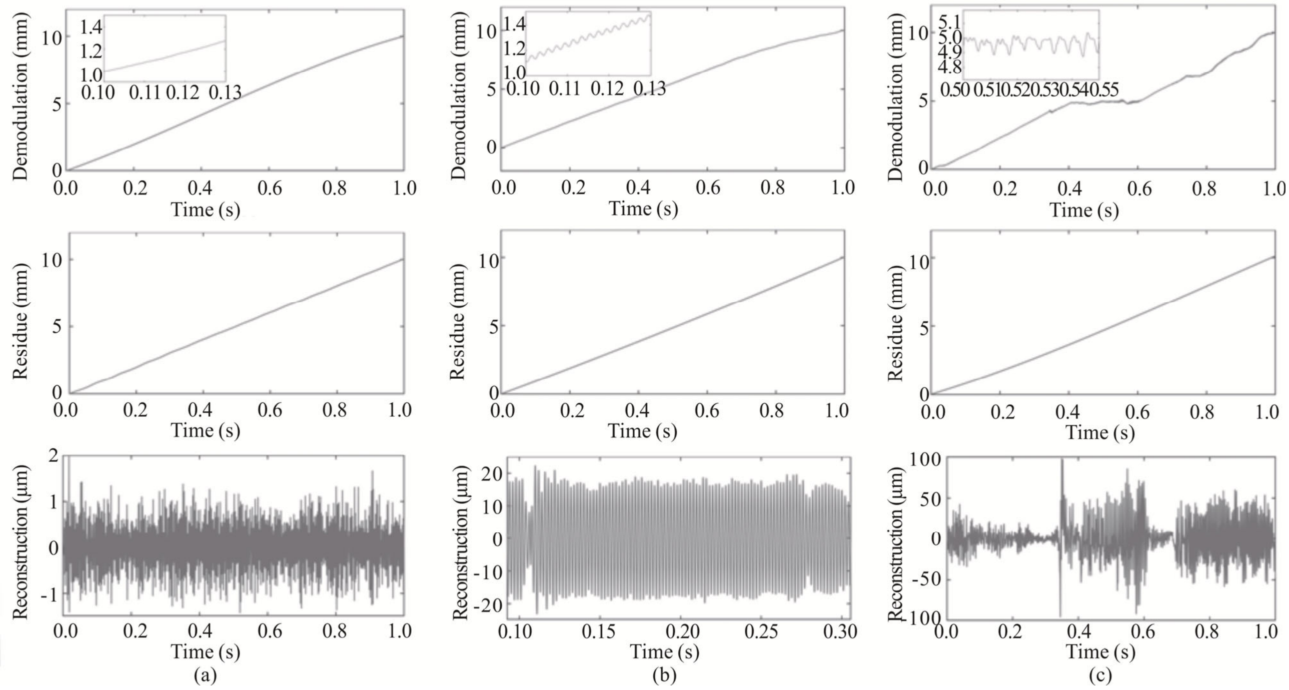

**Fig.3 Demodulated signal, residue (motion component) and reconstructed signal (vibration) of (a) uniform rectilinear movement, (b) uniform rectilinear movement with 500 Hz single-frequency harmonic vibration, and (c) uniform rectilinear movement with speech vibration**

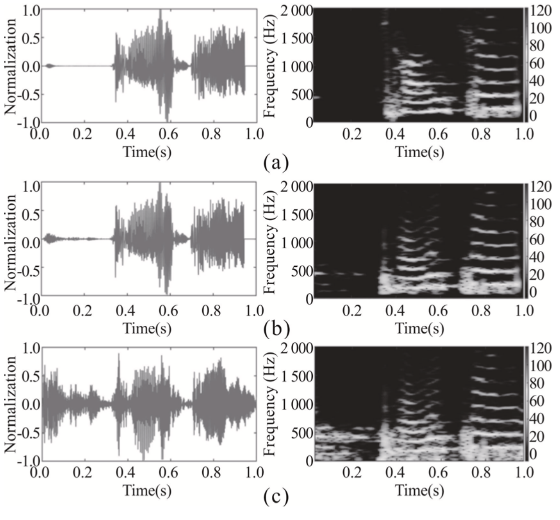

**Fig.4 The speech spectrograms (right) and the normalized amplitude (left) of (a) clean speech, (b) speech enhanced by using OM-LSA on the reconstructed signal, and (c) speech enhanced by using OM-LSA on the overall demodulated signal**

**Tab.1 Results of two different methods**

| Parameter | OM-LSA | EMD & OM-LSA |
|---|---|---|
| *SNR* (dB) | 5.47 | 9.68 |
| Velocity (mm/s) | -- | 10.08 |
| MOS | 3.2 | 3.7 |

This paper associates the EMD algorithm with the OM-LSA algorithm to successfully separate and enhance the speech micro-vibration of a target from the macro motion. The experiments imply that this combined approach has great performance for acquiring and splitting the actual macro motion and micro-vibration characteristics of a target. In addition, compared with the direct enhancement of the demodulated signal by OM-LSA, the enhanced speech signal obtained by this combined algorithm has a higher *SNR* and MOS. The combined algorithm provides a promising, useful approach for speech signal acquisition and enhancement under macro motion.

## References

[1] Sugimoto T, Sugimoto K, Uechi I, Ohdaira T, Kawakami A and Utagawa N, Long Distance Measurement over 30m by High-speed Noncontact Acoustic Inspection Method Using Acoustic Irradiation Induced Vibration, 2017 IEEE International Ultrasonics Symposium (IUS), 1 (2017).

[2] Yun-Peng Wang, Yi-Hua HU, Li-Ren Guo and Shi-Long Xu, Acta Photonica Sinica **7**, 135 (2017). (in Chinese)

[3] Kurvinen E, John M and Mikkola A, Measurement **150**, 107091 (2020).

[4] Frehlich R, Hannon S M and Henderson S W, Boundary-Layer Meteorology **86**, 233 (1998).

[5] Chong Wang, Haiyun Xia, Yanping Liu, Shengfu Lin and Xiankang Dou, Optics Communications **424**, 48 (2018).

[6] Ning Xiang and Sabatier J M, IEEE Geoscience and Remote Sensing Letters **1**, 292 (2004).

[7] Aranchuk V, Lal A K, Hess C F and Sabatier J M, Optical Engineering **45**, 104302 (2006).

[8] Kaplan A D, OrSullivan J A, Sirevaag E J, Lai P and Rohrbaugh J W, IEEE Transactions on Biomedical Engineering **59**, 744 (2012).

[9] Yi J, Liu W, Chen S, Backman V, Sheibani N, Sorenson C M, Fawzi A A, Linsenmeier R A and Zhang H F, Light: Science & Applications **4**, 334 (2015).

[10] Wissmeyer G, Pleitez M A, Rosenthal A and Ntziachristos V, Light: Science & Applications **7**, 53 (2018).

[11] Avargel Y and Cohen I, 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, 109 (2011).

[12] Lü Tao, Guo Jin, Zhang He-yong, Yan Chun-hui and Wang Can-jin, Optoelectronics Letters **13**, 275 (2017).

[13] Lü Tao, Han Xiyu, Wu Shisong and Li yuanyang, Optics Communications **440**, 117 (2019).

[14] Chun-Hui Yan, Ting-Feng Wang, Yuan-Yang Li, Tao Lv and Shi-Song Wu, Chinese Physics B **28**, 030701 (2019).

[15] Y. Deng, 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), 1 (2016).

[16] Peng Shuping, Lü Tao, Wu Shisong, Yan Chunhui and Zhang Heyong, Applied Acoustics **143**, 165 (2019).

[17] Shisong Wu, Yuanyang Li, Tao Lü, Hongkai Chen, Chunhui Yan, Tingfeng Wang and Jin Guo, Chinese Optics Letters **17**, 051201 (2019).

[18] Huang N E, Shen Z, Long S R, Wu M C, Shih H H, Zheng Q, Yen N C, Tung C C and Liu H H, Proceedings A **454**, 903 (1998).

[19] Cohen I and Berdugo B, Signal Processing **81**, 2403 (2001).