MFCC based real-time speech reproduction and recognition using distributed acoustic sensing technology*

ZHOU Ran[1,2,3], ZHAO Shuai[1,2,3], LUO Mingming[1,2,3]**, MENG Xin[1,2,3], MA Jie[1,2,3], and LIU Jianfei[1,2,3]

1. *School of Electronic Information Engineering, Hebei University of Technology, Tianjin 300401, China*
2. *Tianjin Key Laboratory of Electronic Materials and Devices, Hebei University of Technology, Tianjin 300401, China*
3. *Hebei Key Laboratory of Advanced Laser Technology and Equipment, Hebei University of Technology, Tianjin 300401, China*

(Received 18 August 2023; Revised 13 October 2023)

The distributed acoustic sensing technology was used for real-time speech reproduction and recognition, in which the voiceprint can be extracted by the Mel frequency cepstral coefficient (MFCC) method. A classic ancient Chinese poem “You Zi Yin”, also called “A Traveler’s Song”, was analyzed both in time and frequency domains, where its real-time reproduction was achieved with a 116.91 ms time delay. The smaller scaled MFCC$_0$ at 1/12 of MFCC matrix was taken as a feature vector of each line against the ambient noise, which provides a recognition method via cross-correlation among the 6 original and recovered verse pairs. The averaged cross-correlation coefficient of the matching pairs is calculated to be 0.580 6 higher than 0.188 3 of the nonmatched pairs, promising an accurate and fast method for real-time speech reproduction and recognition over a passive optical fiber.

Document code: A **Article ID:** 1673-1905(2024)04-0222-6

DOI https://doi.org/10.1007/s11801-024-3167-5

As an indispensable information carrier, sound records the inheritance of culture and memory for mankind. The sound is always received, distinguished and stored in the form of magnetism in tape, electricity in circuit and light in discs[1-3]. Compared with electroacoustic transducer, passive acoustic sensors based on optical fiber stand out for distant monitoring, compatible installation, and immunity to electromagnetics[4-6]. The central wavelength fluctuation is regarded as an indicator to vibration in fiber Bragg grating (FBG) sensors[7], while the multiplexing is hard to achieve for its limited cascade and bandwidth. The probe arm in Mach-Zehnder interferometer (MZI) is used for vibration sensing as well[8]. Nevertheless, the vibration source is hard to locate due to the phase accumulation along the sensing path, especially in the loop with multiple vibration sources. Thus, the distributed sensing technique based on phase-sensitive optical time domain reflectometry (Φ-OTDR) was employed for qualitative vibration monitoring with backward Rayleigh scattering (RS)[9], which was also termed as the distributed vibration sensor (DVS).

With further investigation on solutions including phase generated carrier (PGC)[10], I/Q demodulation[11], and unwrapping in spatial/time domain[12], distributed acoustic sensing (DAS) technology is optimized with stronger computability[13]. Through the latest improvements in optical system[14,15], photoelectric detection[16], and demodulation algorithms[17], the DAS is close to engineering used for information extraction as a distributed microphone[18,19]. Considering the enormous data stream in acquisition, the reproduction and recognition of human voice is usually processed off-line[20]. Subject to the surroundings laid with cables in field-test, the traditional schemes for signal processing face challenges in locating and recognizing disturbances accurately. So, lots of useful feature extraction and classifier design work has been involved to enhance the perception and recognition ability of DAS[21-23]. An intelligent target recognition method by utilizing both manual and deep features is proposed to make full use of the effective information carried by DAS signals[24], but this method rely heavily on expert knowledge to build system. Then, versatile methods in image processing and speech recognition were transferred to DAS, where Mel frequency cepstral coefficient (MFCC) feature and convolutional neural network (CNN) were applied for classification in vibration behavior[25,26]. A complex convolution recurrent network (CCRN) algorithm is constructed to enhance the information identification of speech signals[27]. However, the training execution time is exceptionally long because

* This work has been supported by the National Natural Science Foundation of China (No.6210031560), the Natural Science Foundation of Hebei Province (No.A2020202013), and the Natural Science Foundation of Tianjin City (No.21JCQNJC00780).

** E-mail: mmluo@hebut.edu.cn

the extensive dataset and the model's architecture are complex, and efficient central processing unit (CPU) and graphics processing unit (GPU) methods are required to improve efficiency[28]. Moreover, it can be difficult to tune the structure and parameters of the network for optimal performance. Therefore, vectors with sufficient characteristics and smaller capacity are necessary in fast voiceprint reproduction and recognition. It is crucial to be able to rapidly and efficiently recognize signals in real time while avoiding increased complexity.

In this paper, MFCC_0 is extracted as a feature vector of voiceprint for real-time speech reproduction and recognition. An ancient Chinese poem of 6 lines, "A Traveler's Song", was analyzed by MFCC method and reproduced with 116.91 ms time delay. With the correlation with original data as an indicator, the 618×1 MFCC_0 is proved noise-resist in numerical analysis. In experiment, the correlation in MFCC_0 between each original verse and each recovered verse was given compared respectively. The averaged correlation of the matching verse pairs is 0.580 6 higher than 0.188 3 of the non-matched pairs, which promises an accurate and time-saving method for voiceprint recognition.

Defined as the characteristic parameters extracted from sound spectrum based on human auditory mechanism[29], MFCC accurately describes the nonlinear characteristics of the human ear and the basic features of sound. The scaling of perceiving frequency is linear up to 1 kHz and above is logarithmic[30], thus, the mapping between linear frequency to Mel-frequency is expressed as[31]

$$F_{\mathrm{mel}} = k_{\mathrm{const}} \cdot \log_{10}(1 + f / f_0), \quad (1)$$

where the transfer constant k_{const} is set as 2 595 and frequency threshold of human sound f_0 is 700, which makes feature extraction fit for human auditory perception. A higher information density and a smaller file capacity become available in MFCC matrix when compared with direct analysis in time domain and frequency domain. Therefore, the extraction of MFCC features promises an effective approach to real-time speech recognition and sound retrieval.

Three steps comprise the extraction procedure of MFCC from the speech sample as the dashed frames shown in Fig.1, including the pre-processing, conversion to Mel-frequency, and cepstrum analysis. The detailed processes are further expressed below.

(1) For the intensity attenuation at higher frequency range through the transmission, pre-emphasis with a high-pass filter eliminates the effects of vocal cords and lips during vocalization:

$$H(z) = 1 - \mu z^{-1}, \quad (2)$$

where μ is between 0.9 and 1, usually 0.97.

(2) The signal is divided into several frames of 20—30 ms with an 75% overlap between the adjacent segments.

(3) The Hamming window is applied upon each frame respectively, to eliminate discontinuity between adjacent ones, which has the form as

$$W(n,a) = (1-a) - a \times \cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1, \quad (3)$$

where the value of a is 0.46.

(4) The windowed signal is expanded frame by frame in frequency domain with fast Fourier transform (FFT) algorithm as

$$X(i,k) = \mathrm{FFT}(x_i(n)). \quad (4)$$

(5) The signal is filtered by the Mel filter bank, and the filter response $H_m(k)$ is defined as

$$H_m(k) = \begin{cases} 0, k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, f(m) < k \le f(m+1) \\ 0, k > f(m+1) \end{cases}, \quad (5)$$

where $1 \leqslant m \leqslant M$, and the number of filters M is set as 24.

(6) The logarithmic energy output is evaluated for each filter as

$$s(i,m) = \log\left(\sum_{k=0}^{N-1} |X(i,k)|^2 H_m(k)\right). \quad (6)$$

(7) The high-dimensional MFCC matrix is obtained through the discrete cosine transform (DCT) as

$$\mathrm{MFCC}(i,n) = \sum_{m=0}^{M-1} s(i,m) \cos\left(\frac{\pi n(m-0.5)}{M}\right), \quad (7)$$

where $0 \leqslant n \leqslant L-1$, and L is the desired number of cepstral features.

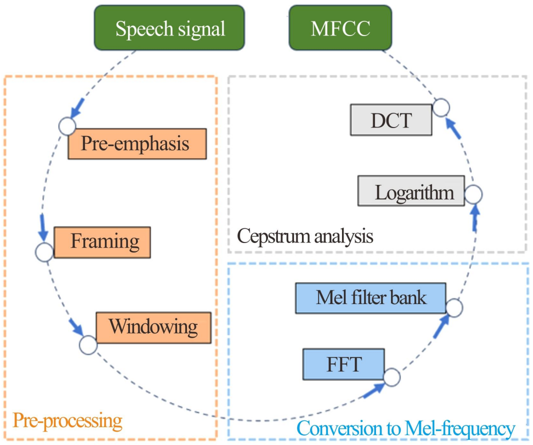

Fig.1 Schematic logic diagram of MFCC feature extraction

Then, the first value of MFCC of each frame is taken to form a new feature vector MFCC_0, for its outstanding representativeness and immunity to different tones when compared with the other dimensions. Moreover, the MFCC_0 vector has a much smaller file size involved in operation than the whole matrix, not to mention the uncompressed raw data of the speech.

For an audio signal with a duration of 4 s and a sampling rate of 20 kHz, as an example, the frame capacity is set to 512 points (25.6 ms/frame) to ensure the robustness inside the frame. Besides, 384 points (75% overlap) are utilized to maintain the continuity between two adjacent frames. The normalized MFCC$_0$ becomes a 618×1 feature vector extracted from 618×12 MFCC matrix, which indicates characteristics of the 80 000 points original signal. In numerical analysis, the Gaussian white noise was performed to test noise-resistance of the MFCC$_0$ scheme. In Fig.2(a), the original audio signal and the one with a noise of 15 dB were normalized and depicted with blue and red lines, respectively. The MFCC$_0$ spectra of audio signals with and without noise were extracted and traced with corresponding colors in Fig.2(b). The feature lines change with the trend of the audio signal, where the larger amplitude appears at the fluctuation with higher frequency components. Despite the attached Gaussian white noise, MFCC$_0$ curve of the signal with noise is proved basically consistent with that of the original one.

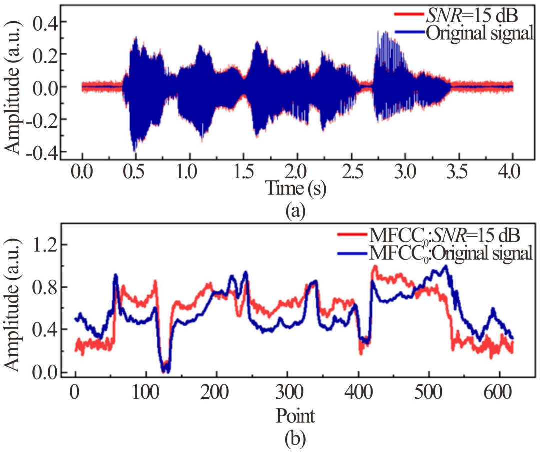

Fig.2 (a) Contrast in time domain signals and (b) their MFCC$_0$ vectors from numerical analysis

Quality of the audio is improved with the increasing signal-to-noise ratio (*SNR*), which makes real-time speech recognition complicated and difficult at a low *SNR*. In Fig.3, the blue scatters refer to the cross-correlation in MFCC$_0$ among the original audio and those with additional noise from 0 dB to 100 dB. The polynomial fitting curve increases to saturation 1 at 58.06 dB, where it is 0.648 4 at 15 dB. The cross-correlation in MFCC$_0$ of the recovered signal and the original signal is used to measure the correlation between the two for identification. When the *SNR* is reduced to 10 dB, the correlation coefficient is 0.583 1, indicating that the correlation is medium and can still be recognized. The analysis shows that the MFCC$_0$ can be regarded as an indicator for real-time speech recognition.

The schematic illustration of DAS system using coherent detection in experiment is shown in Fig.4. The 1 550.12 nm light from a 3 kHz narrow linewidth laser (NLL) propagates through a 90: 10 coupler (OC1), where the 90% of the 20 mW optical power goes to an acousto-optic modulator (AOM) with 80 MHz shift. The pulsed 100 ns light was then injected into the fiber under test (FUT) via a circulator (CIR) after being amplified by an erbium doped fiber amplifier (EDFA). The backward RS interferes with the rest 10% reference light at a 50: 50 coupler (OC2), which is harvested by a balanced photo detector (BPD) and collected by a dual-channel digital data acquisition (DAQ) card with 250 Msps and 14-bit A/D for post-processing. The scanning rate was set to 20 kHz, where the total number of frames exceeds 80 000 in 4 s sampling time. Loudspeaker 1 plays audio through Bluetooth as a sound source placed next to the FUT. The source was placed at 800 m away from the DAS system along the 5 km sensing fiber, where 2 m non-armored fiber was used for speech reproduction and recognition with digital quadrature demodulation. The loudspeaker 2 is connected to computer processing terminal to play the recovered audio demodulated from the DAS system.

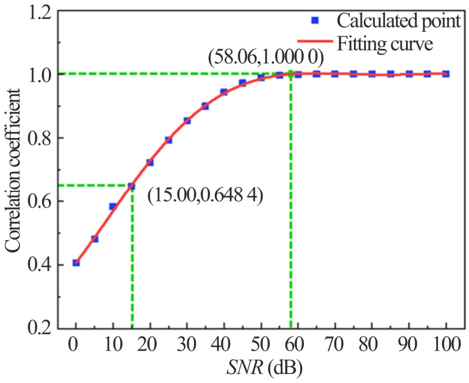

Fig.3 *SNR* dependent correlation in MFCC$_0$ vector with its polynomial fitting curve

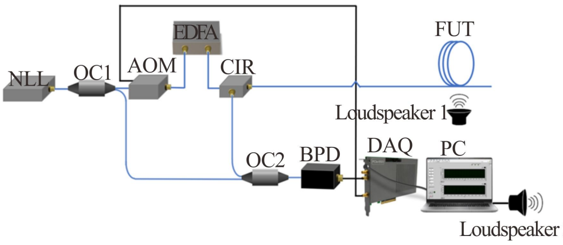

Fig.4 Schematic setup of the Φ-OTDR based DAS system

The 6 verses of an ancient Chinese poem "A Traveler's Song" were recorded by the DAS and recovered in Fig.5. The original and recovered voiceprints were displayed with blue and red lines, respectively, where the 4 s duration for each verse was normalized between −1 and 1. Though differences in detail were observed between original and recovered signals, their unique outlines and characteristics are identifiable still. Moreover, the recovered verses can be heard and recognized just by human ears referring to attached supplementary video materials. And the stronger the speech signal, the clearer the reproduced speech signal. When the strength of the speech signal is reduced to 80 dBA, there is more noise in the recovered signal, which reduces the clarity, but it can still be recognized.

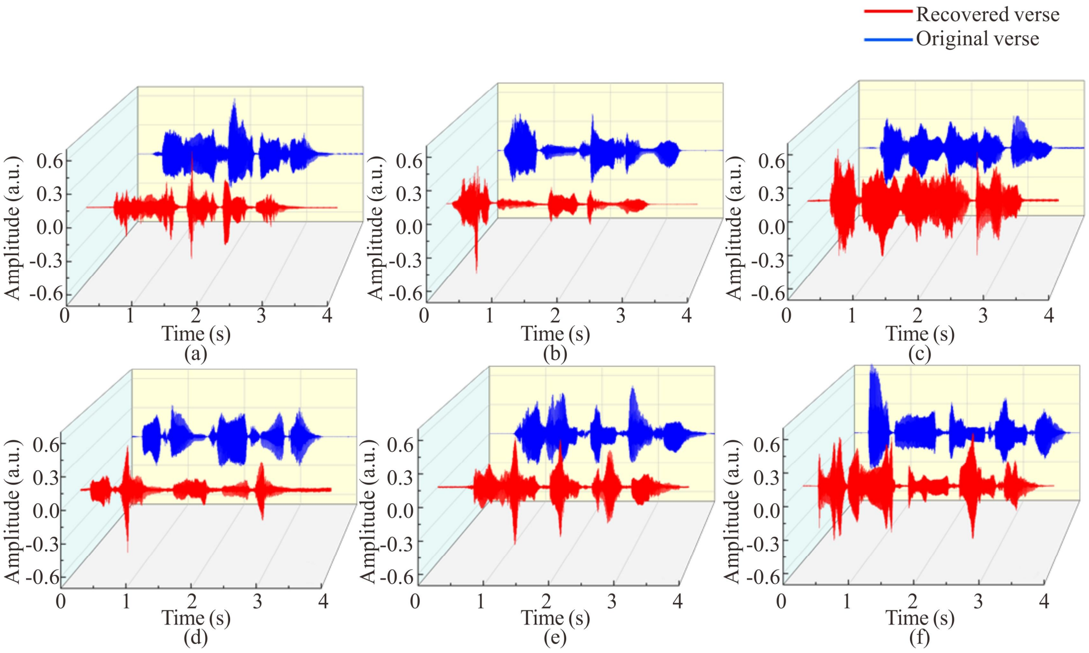

Fig.5 The original and recovered audio signals of all the 6 lines in experiment: (a) Verse 1; (b) Verse 2; (c) Verse 3; (d) Verse 4; (e) Verse 5; (f) Verse 6

The third verse "Lin Xing Mi Mi Feng" ("Sewn stitch by stitch before he leaves"), as an example, was analyzed in frequency spectrum. By direct FFT processing with additional noise, the frequency components of the original and recovered verses contrastively present in Fig.6. Their distributions are similar below the threshold of human voice at 700 Hz aforementioned, with a same strongest peak at 208 Hz. The frequency range in which the DAS detects speech signals is also tested, and as the frequency increases beyond 700 Hz, there is a large attenuation of the amplitude, which affects the performance of the signal reproduction. This is demonstrated by the fact that at higher frequency beyond 700 Hz in the red dashed frame, the intensity was suppressed for its lower response to high frequency. However, the absence of high frequency components is not a big deal because the main frequency distribution remains still.

The MFCC$_0$ for original and recovered verse 3 were extracted from the MFCC matrix derived from signals in Fig.7(a) separately. As the blue and red lines depicted in Fig.7(b), the original and recovered MFCC$_0$ vectors are similar with each other and even consistent with numerical simulation in Fig.2(b). Moreover, the cross-correlation in MFCC$_0$ was calculated to be 0.590 7, which indicates a high similarity between original and recovered verses. Meanwhile the *SNR* in experiment is calculated to be 15.20 dB, which corresponds to cross-correlation of 0.654 3 according to the curve in Fig.3. The actual correlation is lower than the estimated value using Gaussian white noise only, where coeffects appear with different kinds of noises including shot noise, quantum noise, and spontaneous emission noise[32]. The optical fiber is noise-free as a passive element, though the external vibration noise easily enters the FUT by direct contact. Accordingly, all the causes above mentioned may lead to decrease in correlation.

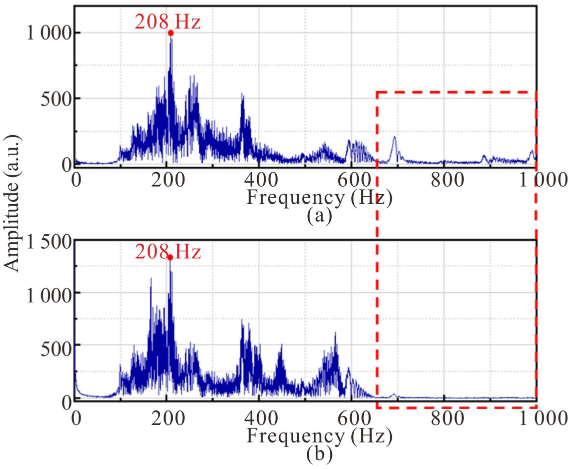

Fig.6 Distributions in frequency domain of the (a) original and (b) recovered verse 3 as an example

The correlation in MFCC$_0$ is calculated for arbitrary pair between the original 6 verses and recovered 6 verses, respectively. In Fig.8(a), each broken line in different color represents the correlation between a certain recovered verse and the original verse 1 to 6. The averaged correlation for the 6 matching verse pairs is 0.580 6, while that of the nonmatching verse pairs is 0.188 3. The highest point in correlation presents at the place where the recovered verse matches its original one. Therefore, the speech can be recognized by searching the similar voice in database. Experimental results are further demonstrated with a 3-dimensional diagram in Fig.8(b), where the highest correlation is lined along diagonal of the grid.

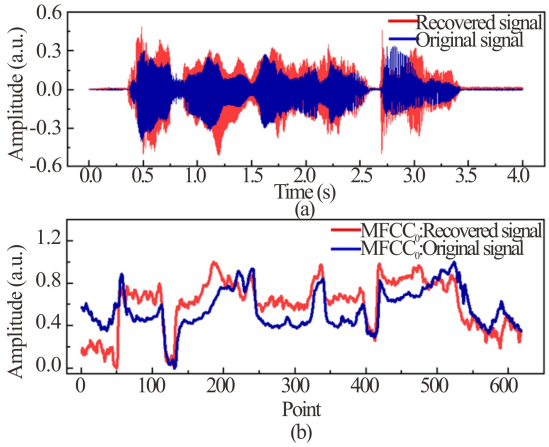

Fig.7 (a) Contrast in time domain signals and (b) their MFCC$_0$ vectors from experimental results

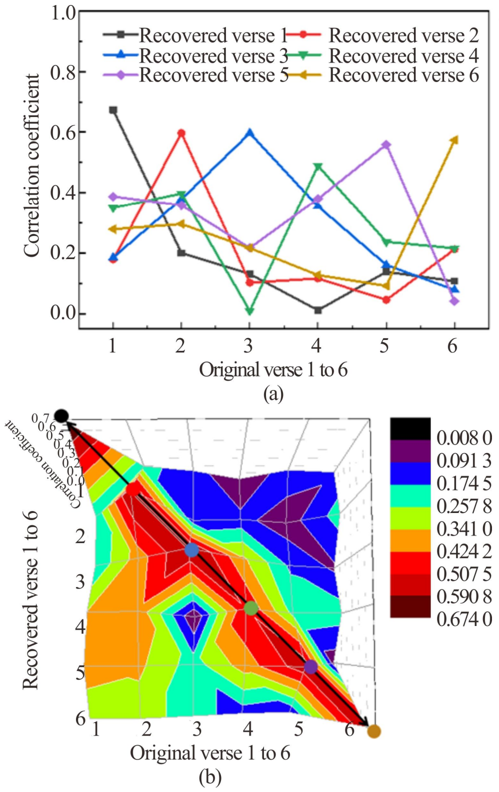

Fig.8 Correlation in MFCC$_0$ among the original and recovered verses with a ridge along the matching pairs: (a) 2-dimensional diagram; (b) 3-dimensional diagram

A small scaled feature vector MFCC$_0$ is demonstrated as an indicator for real-time speech reproduction and recognition using DAS technology. In experiment, the 6 verses of an ancient Chinese poem can be real-time reproduced with 116.91 ms time delay at 1 kHz frame rate. The averaged correlation coefficient was calculated to be 0.580 6 of the 6 matched verse pairs over 0.188 3 of the nonmatching pairs, which quantitatively ensures the recovered audio recognizable among the verses. Our demonstration promises an accurate and fast method for real-time speech reproduction and recognition using DAS, not just for human speech but for ordinary sounds like the pipeline leakage, auto oscillation in power cable, and resonance in architecture.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] WITTJE R. The electrical imagination: sound analogies, equivalent circuits, and the rise of electroacoustics, 1863-1939[J]. Osiris, 2013, 28(1): 40-63.

[2] HUANG Y, ZHOU X. Non-reciprocal sound transmission in electro-acoustic systems with time-modulated circuits[J]. Acta mechanica solida sinica, 2022, 35(6): 940-948.

[3] SHERIF M M, KHAKIMOVA E M, TANKS J, et al. Cyclic flexural behavior of hybrid SMA/steel fiber reinforced concrete analyzed by optical and acoustic techniques[J]. Composite structures, 2018, 201: 248-260.

[4] JOE H E, YUN H, JO S H, et al. A review on optical fiber sensors for environmental monitoring[J]. International journal of precision engineering and manufacturing-green technology, 2018, 5(1): 173-191.

[5] HUBBARD P G, XU J, ZHANG S, et al. Dynamic structural health monitoring of a model wind turbine tower using distributed acoustic sensing (DAS)[J]. Journal of civil structural health monitoring, 2021, 11(3): 833-849.

[6] FERNANDZE-RUIZ M R, SOTO M A, WILLIAMS E F, et al. Distributed acoustic sensing for seismic activity monitoring[J]. APL photonics, 2020, 5(3).

[7] MOCCIA M, PISCO M, CUTOLO A, et al. Opto-acoustic behavior of coated fiber Bragg gratings[J]. Optics express, 2011, 19(20): 18842-18860.

[8] MACIA-SANAHUJA C, LAMELA H, GARCIA-SOUTO J A. Fiber optic interferometric sensor for acoustic detection of partial discharges[J]. Journal of optical technology, 2007, 74(2): 122-126.

[9] XIONG J, WANG Z, JIANG J, et al. High sensitivity and large measurable range distributed acoustic sensing with Rayleigh-enhanced fiber[J]. Optics letters, 2021, 46(11): 2569-2572.

[10] FANG G, XU T, FENF S, et al. Phase-sensitive optical time domain reflectometer based on phase-generated carrier algorithm[J]. Journal of lightwave technology, 2015, 33(13): 2811-2816.

[11] WANG Z, ZHANG L, WANG S, et al. Coherent Φ-OTDR based on I/Q demodulation and homodyne detection[J]. Optics express, 2016, 24(2): 853-858.

[12] TU G, ZHANG X, ZHANG Y, et al. The development of an Φ-OTDR system for quantitative vibration measurement[J]. IEEE photonics technology letters, 2015, 27(12): 1349-1352.

[13] WANG S, JIANG J, WANG S, et al. GPU-based fast processing for a distributed acoustic sensor using an LFM pulse[J]. Applied optics, 2020, 59(35): 11098-11103.

[14] ZHU K, ZHOU B, WU H, et al. Multipath distributed acoustic sensing system based on phase-sensitive optical time-domain reflectometry with frequency division multiplexing technique[J]. Optics and lasers in engineering, 2021, 142: 106593.

[15] ZHANG X, QIAO W, SUN Z, et al. A distributed optical fiber sensing system for synchronous vibration and loss measurement[J]. Optoelectronics letters, 2016, 12(5): 375-378.

[16] LU Y, ZHU T, CHEN L, et al. Distributed vibration sensor based on coherent detection of phase-OTDR[J]. Journal of lightwave technology, 2010, 28(22): 3243-3249.

[17] DONG Y, CHEN X, LIU E, et al. Quantitative measurement of dynamic nanostrain based on a phase-sensitive optical time domain reflectometer[J]. Applied optics, 2016, 55(28): 7810-7815.

[18] MASOUDI A, BELAL M, NEWSON T P. Distributed optical fiber audible frequency sensor[C]//23rd International Conference on Optical Fiber Sensors, June 2-6, 2014, Santander, Spain. Washington: SPIE, 2014: 537-540.

[19] FRANCISCANGELIS C, MARGULIS W, KJELLBERG L, et al. Real-time distributed fiber microphone based on phase-OTDR[J]. Optics express, 2016, 24(26): 29597-29602.

[20] WU Y, GAN J, LI Q, et al. Distributed fiber voice sensor based on phase-sensitive optical time-domain reflectometry[J]. IEEE photonics journal, 2015, 7(6): 1-10.

[21] ZHANG P, VENKETESWARAN A, WRIGHT R, et al. Feature extraction for pipeline defects inspection based upon distributed acoustic fiber optic sensing data[C]//Fiber Optic Sensors and Applications XVIII, April 3-June 12, 2022, Virtual. Washington: SPIE, 2022: 14-29.

[22] TABJULA J, SHARMA J. Feature extraction techniques for noisy distributed acoustic sensor data acquired in a wellbore[J]. Applied optics, 2023, 62(16): E51-E61.

[23] NING F, CHENG Z, MENG D, et al. A framework combining acoustic features extraction method and random forest algorithm for gas pipeline leak detection and classification[J]. Applied acoustics, 2021, 182: 108255.

[24] WU H, WANG C, LIU X, et al. Intelligent target recognition for distributed acoustic sensors by using both manual and deep features[J]. Applied optics, 2021, 60(23): 6878-6887.

[25] JIANG F, LI H, ZHANG Z, et al. An event recognition method for fiber distributed acoustic sensing systems based on the combination of MFCC and CNN[C]//2017 International Conference on Optical Instruments and Technology: Advanced Optical Sensors and Applications, October 28-30, 2017, Beijing, China. Washington: SPIE, 2018: 15-21.

[26] SHI Y, LIU X, WEI C. An event recognition method based on MFCC, superposition algorithm and deep learning for buried distributed optical fiber sensors[J]. Optics communications, 2022, 522: 128647.

[27] SHANG Y, YANG J, CHEN W, et al. Speech signal enhancement based on deep learning in distributed acoustic sensing[J]. Optics express, 2023, 31(3): 4067-4079.

[28] BENCHARIF B A E, BAYAR S, ÖZKAN E. Parallel implementation of distributed acoustic sensor acquired signals: detection, processing, and classification[J]. Journal of applied remote sensing, 2022, 16(2): 024504-024504.

[29] AYVAZ U, GURULER H, KHAN F, et al. Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning[J]. CMC-computers materials & continua, 2022, 71(3).

[30] ARPITHA Y, MASHUMATHI G L, BALAJI N. Spectrogram analysis of ECG signal and classification efficiency using MFCC feature extraction technique[J]. Journal of ambient intelligence and humanized computing, 2022, 13(2): 757-767.

[31] GANCHEV T, FAKOTAKIS N, KOKKINAKIS G. Comparative evaluation of various MFCC implementations on the speaker verification task[C]//Proceedings of the SPECOM, October 17-19, 2005, Patras, Greece. Moscow, 2005: 191-194.

[32] BLOTEKJAER K. Fundamental noise sources that limit the ultimate resolution of fiber optic sensors[C]//Optical and Fiber Optic Sensor Systems, September 16-19, 1998, Beijing, China. Washington: SPIE, 1998: 1-12.