
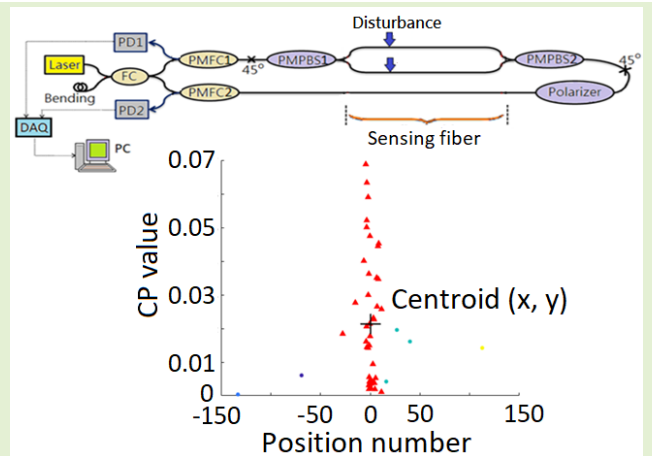


Dual Mach–Zehnder Interferometers With Hierarchical Clustering Analysis Method Applied for Positioning the Intrusion

Meng-Chen Li and Likarn Wang 

Abstract—The method of hierarchical clustering is for the first time employed to locate the intrusion-induced disturbance on the sensing fibers of a dual Mach–Zehnder interferometer (DMZI). Such an intrusion-induced disturbance is located by finding the x coordinate of the centroid of the largest cluster on the Euclidean plane through hierarchical clustering with an appropriate linkage criterion employed for determining the distance between two observations. We compare average linkage and complete linkage criteria in the clustering analysis to see which one provides better locating accuracy. In the clustering analysis, the number of clusters is set to be 3–8 in finding the location of disturbance. To reduce the locating error, we also use differential signals here in the clustering analysis. Twelve intrusion events are simulated by knocking the sensing fibers to induce disturbances at a given location. The location of disturbance is determined through the clustering analysis for each intrusion event. The mean of the absolute values of locating errors [mean absolute error (MAE)] for the 12 intrusion events is then estimated. The experimental results in this study demonstrate a maximum MAE of 11.55 m in locating an intrusion with average linkage criterion employed for five-cluster analysis. Also, the MAE could be 3.55 m smaller by using the differential signals for clustering analysis, compared with the case when directly detected signals are used for clustering analysis. The results also confirm that the average linkage criterion provides only a small amount of improvement in MAE over complete linkage criterion.

Index Terms—Average linkage criterion, complete linkage criterion, dual Mach–Zehnder interferometer (DMZI), fiber-optic intrusion detection, hierarchical clustering, positioning accuracy.



I. INTRODUCTION

TWO kinds of fiber intrusion detection techniques have been widely studied for detecting and locating the disturbance on the sensing fiber. One kind is to detect and locate the disturbance-induced phase variation in Rayleigh backscattered light employing the optical time-domain interferometer-based technique [1], [2], [3], [4], [5].

Meanwhile, optical interferometers have also been studied for the purpose of intrusion detection. A Sagnac loop

interferometer was proposed for positioning of a burst acoustic wave along a 35-km fiber loop to be within tens of meters [6]. A disturbance sensor scheme with two time-division multiplexed Sagnac interferometers achieved a maximum positioning error of 400 m in a test of a 10-km-long fiber [7]. Michelson interferometers were also used for distributed disturbance detection [8], [9]. Systems that combined different types of interferometers were also studied for detecting and locating the disturbance applied on sensing fibers, such as those with merged Sagnac and Michelson interferometers [10], [11] and those with combined Sagnac and Mach–Zehnder interferometers [12], [13].

Recently, many research groups have paid attention to dual Mach–Zehnder interferometers (DMZIs) for disturbance detection. In a traditional DMZI system, a laser light is split into two paths for a clockwise (CW) and a counterclockwise (CCW) interferometer. To determine the time delay between the two signals detected by the CW and the CCW interferometers, which corresponds to the location of disturbance, a cross correlation algorithm is usually used. In the experiment of [14], an average locating error of 390 m for an 18.46-km-long detection range was obtained. In a work using a polarization

Manuscript received 23 September 2022; revised 15 November 2022; accepted 24 November 2022. Date of publication 9 December 2022; date of current version 12 January 2023. This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST111-2221-E-007-025. The associate editor coordinating the review of this article and approving it for publication was Dr. Rajan Jha. (Corresponding author: Likarn Wang.)

Meng-Chen Li was with the Institute of Photonics Technologies, National Tsing Hua University, Hsinchu 30013, Taiwan. He is now with Taiwan Semiconductor Manufacturing Company, Hsinchu 300-096, Taiwan (e-mail: lemon111276@gmail.com).

Likarn Wang was with the Institute of Photonics Technologies, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: lkwang@ee.nthu.edu.tw).

Digital Object Identifier 10.1109/JSEN.2022.3226761

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.

For more information, see <https://creativecommons.org/licenses/by-nc-nd/4.0/>

control method to eliminate the effect of polarization-induced fading (PIF), a locating error of 160 m in a test of 112-km fiber length was obtained [15]. A chaotic particle swarm optimization algorithm was used for eliminating the PIF effect in the work of [16], and a locating error of ± 20 m over a 2.25-km sensing cable was reported, which was much lower than that obtained by using the traditional polarization control method based on the criterion of interference visibility. The technique of wavelength-division multiplexing was proposed to reduce the influence of Rayleigh backscattering on the signals detected in the DMZIs. A locating error of 52.5 m was reported for the case of 61-km sensing length, which was much better than that with traditional DMZIs [17]. Faraday rotating mirrors were employed in the work of [18] for eliminating the effect of PIF in the experiment of 100-km sensing distance with a locating error of ± 25 m.

We have proposed a Fourier spectral analysis (FSA) method to determine the location of disturbance for a modified DMZI system [19]. In the FSA method, the spectrally dependent locations of disturbance were calculated and the average of the locations over an appropriate spectral band gave the location of disturbance. A long-term test has demonstrated the reliability of this system, which does not use any polarization control method to eliminate the PIF effect. For a fiber length of 250 m, a maximum locating error of 26 m was obtained in a test comprising five intrusion events occurring at a given position. To further enhance the positioning accuracy, here, we apply a hierarchical clustering analysis method in determining the location of disturbance. In Section II, we briefly review the modified DMZIs system and describe the method of hierarchical clustering analysis, where we show the clustering results obtained by using the differential signals that are obtained by taking difference operation on the directly detected signals and, then, we compare the results with those obtained by using the directly detected signals for a given case of intrusion. Section III gives an estimation of locating error for an intrusion event that lasts for 1 s and shows a reliability test result by disturbing the sensing fibers 12 times at a given position. Then, a conclusion is given in Section IV.

II. OUTLINE OF THE PRESENTED SYSTEM

A. Experiments

The DMZI system used in the study is as same as that in [19] and is shown in Fig. 1, where a 1036-m-long fiber cable comprising four single-mode fibers (SMFs) is used as a disturbance sensor. A polarization-maintaining fiber coupler (PMFC1) is connected to a polarization-maintaining polarization beam splitter (PMPBS1) with their principal axes oriented at 45° relative to each other. At the other end of the DMZI, a polarization-maintaining polarization beam splitter (PMPBS2) is connected with a polarization-maintaining fiber polarizer (Polarizer) at 45° between the principal axes of the two components. A laser light is first split into a CW-propagating and a CCW-propagating light, which are referred to, respectively, as CW and CCW lights, through a fiber coupler (FC). The CW light splits into two orthogonally polarized lights (say x -polarized and y -polarized lights) at the two sensing arms

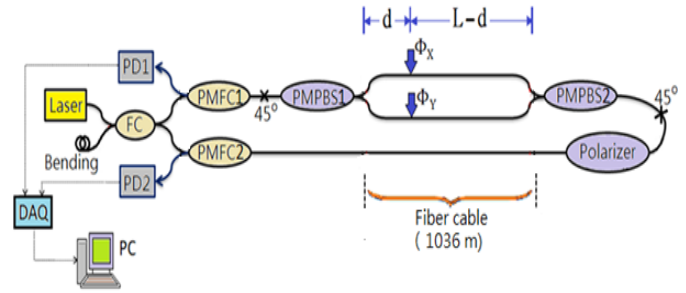


Fig. 1. DMZI sensor system used in this study. As the sensing arms are disturbed at a location of d from PMPBS1, phase variations of light in the two sensing arms, Φ_x and Φ_y , are generated.

of the DMZI. Meanwhile, the CCW light also splits into two orthogonally polarized waves at the two sensing arms after passing through polarizer and PMPBS2.

As the sensing fibers are disturbed at a distance of d from PMPBS1, phase variations of light in the two sensing fibers, Φ_x and Φ_y , are generated, and the optical powers detected by photodetectors PD1 and PD2, i.e., $I_{PD1}(t)$ and $I_{PD2}(t)$, can be ideally written as

$$I_{PD1}(t) = C + D \cos(\Phi(t - 2\tau_2)) \quad (1)$$

$$I_{PD2}(t) = A + B \cos(\Phi(t)) \quad (2)$$

respectively, where A – D are constants and Φt is defined as $\Phi_x(t) - \Phi_y(t)$. Equations (1) and (2) reveal a time delay of $2\tau_2 = 2(L - d)/c$ between the detected CW and CCW signals, where c is the light speed in the fiber. The disturbance on the fiber cable could be induced in many ways. For example, it could be induced by intruders who vibrate the fiber cable that is attached on a netted fence. When the fiber cable is vibrated, the phase variations Φ_x and Φ_y are induced due to the strains upon the two fiber arms at the same position. The two detected signals $I_{PD1}(t)$ and $I_{PD2}(t)$ are then transmitted to a personal computer (PC) through a data acquisition module (DAQ), which samples the signal waveforms at a specific sampling rate.

To simulate an intrusion event, we will knock (or heavily tap) the fiber cable at an arbitrarily chosen location and proceed with the calculation detailed in the following to determine the disturbance location for the corresponding intrusion event. The test is also undertaken for intrusion events occurring at another two arbitrarily chosen locations. The derivations for the term $(\Phi_x - \Phi_y)$ appearing in (1) and (2) can be seen from [19, eqs. (2)–(8)]. This term results from the interference between the two waves in the two fiber arms. Note that Φ_x and Φ_y are essentially different due to the different strains imposed on the two fiber arms as the fiber cable is vibrated, and therefore, the term $(\Phi_x - \Phi_y)$ varies upon intrusion.

B. Hierarchical Clustering Method for Locating Disturbance

In the hierarchical clustering method, the Fourier transforms of the signal waveforms $I_{PD1}(t)$ and $I_{PD2}(t)$ within every 10-ms time period are taken separately. The Fourier transforms are

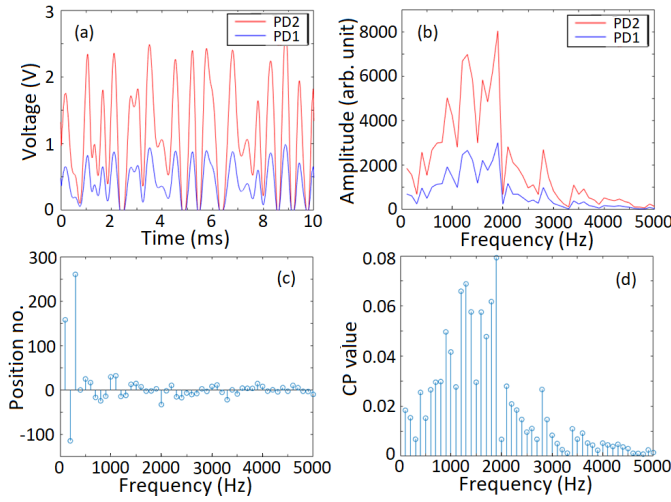


Fig. 2. (a) Signal waveforms received by PD1 and PD2, i.e., $I_{PD1}(t)$ and $I_{PD2}(t)$, within a 10-ms time period; (b) absolute values of their Fourier amplitudes; (c) calculated position number as a function of frequency; and (d) CP value spectrum over the frequency range from 100 to 5000 Hz for an intrusion case when the fiber cable was heavily tapped at the position of 1 m (corresponding to a position number of 0.02) from PMPBS2.

shown in (3) and (4), where $F_{PD1}(\omega)$ and $F_{PD2}(\omega)$ are the corresponding Fourier amplitudes (or spectral amplitude) and ideally are proportional to each other at any ω . Thus, the delay time $2\tau_2$ between I_{PD1} and I_{PD2} at any ω can be found by dividing the phase shift between the Fourier components of $I_{PD1}(t)$ and $I_{PD2}(t)$ by ω . Ideally, the calculated delay time $2\tau_2$ should be

$$I_{PD1}(t) \xleftrightarrow{\text{Fourier transform}} F_{PD1}(\omega) e^{-j(2\tau_2\omega)} \quad (3)$$

$$I_{PD2}(t) \xleftrightarrow{\text{Fourier transform}} F_{PD2}(\omega) \quad (4)$$

constant over the frequency range of concern in this study. However, the delay time varies due to detection noises and nuisances, as revealed in [19, Figs. 6, 7, and 9–14]. This situation can be seen again from an example taken by heavily tapping the fiber cable for simulating an intrusion at the position of $L - d = 1$ m. The position of disturbance corresponds to a position number of 0.02. Here, the position number is defined to be a normalized distance from PMPBS2, i.e., $(L - d)/50$, with L and d both expressed in meter and 50 being the spatial resolution in meter resulting from the sampling rate of the DAQ used in the experiments. Fig. 2 shows for this case the acquired signal waveforms of $I_{PD1}(t)$ and $I_{PD2}(t)$ in Fig. 2(a), absolute values of their Fourier amplitudes in Fig. 2(b), the calculated position number as a function of frequency in Fig. 2(c), and the CP value as a function of frequency over 100–5000 Hz in Fig. 2(d). Here, the characteristic power (CP) value at a given frequency is defined to be a ratio of the spectral amplitude at that frequency to the summation of the spectral amplitudes over the whole frequency band of 100–5000 Hz. Therefore, CP represents the normalized spectral amplitude or power at a given frequency for a detected signal. The CP values at some frequencies, such as 3300, 4600–4800, and 5000 Hz, are relatively low and correspond to weak spectral components. Also, these spectral

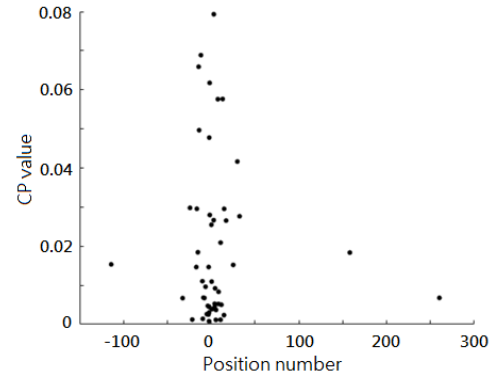


Fig. 3. Two-dimensional feature space filled with 50 data points with position number and CP value being x and y coordinates.

components may be disregarded in determining the location of disturbance because these weak spectral components are susceptible to noises and nuisances.

To apply the clustering method to determine the location of disturbance, a 2-D feature space (i.e., Euclidean space) is first established by using two feature parameters, i.e., position number and CP value, as x and y coordinates, respectively. Fig. 3 shows the 2-D space filled with 50 data points, with each point having x and y coordinates representing, respectively, the position number and CP value taken at a given frequency from 100 to 5000 Hz, as shown in Fig. 2(c) and (d). In this case, all of the data points in the space release the information of strength of the spectral component and the corresponding location of intrusion at a given frequency. These two parameters are related to each other by frequency, as shown in Fig. 2(c) and (d). Because the CP value is typical of the strength of a spectral component, some data points with a low value of y coordinate can be ruled out in determining the location of intrusion. The data points in the feature space can thus be divided into groups using a clustering algorithm to give a dominant group or cluster, the x coordinate of the centroid of which gives the location of intrusion (expressed in position number). K -means clustering and hierarchical clustering methods are commonly used in grouping data samples for various applications, such as data mining and statistics. Either of the clustering methods can be applied to classify the data samples/points into multiple groups, i.e., clusters, and the x coordinate of the centroid of the largest group, i.e., the dominant group, represents the location of intrusion in this case.

Because the clustering result would depend on initially chosen clusters in the K -means clustering algorithm [20], [21], the determined location of intrusion could vary with such initial centroids chosen, thus leading to appreciable uncertainty in locating the disturbance. As can be seen from Figs. 2(c) and 3, several position numbers have quite a large value, which should not be considered in determining the location of intrusion. Clustering analysis is a good method to exclude those data points with large values of position number. Meanwhile, a hierarchical clustering analysis is easy to implement, and one can choose the number of clusters readily from the hierarchical tree called dendrogram [22]. This type

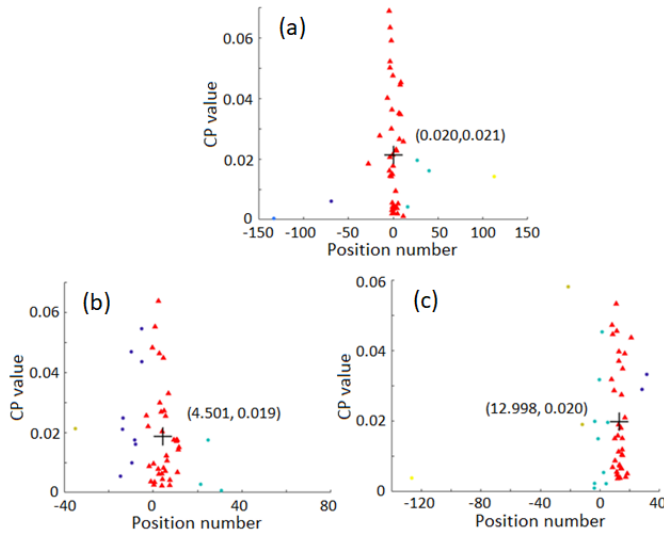


Fig. 4. Five clusters (appearing in different colors) obtained in Euclidean space by using the hierarchical clustering analysis with complete linkage criterion employed for three intrusion events occurring at the positions of (a) $d = 1035$ m, (b) $d = 787$ m, and (c) $d = 393$ m. The centroid of the largest cluster with data points of red triangles is marked by cross with coordinates in each case of the intrusion events.

of clustering is classified as agglomerative clustering (with decomposition of cluster following bottom-up strategy) and divisive clustering (with decomposition of cluster following top-down strategy). Various kinds of linkage criteria can be used for measuring the distance between clusters, such as average linkage criterion and complete linkage criterion. Data samples with similar identities can be classified into the same subgroup, i.e., the same cluster, by employing appropriate linkage criterion [22]. Note that the position numbers at lower frequencies [see Fig. 2(c)] correspond to the data points with large x coordinates, which are classified into smaller groups (see Fig. 3). Because only the largest group will be the focus of attention in determining the location of disturbance, those points belonging to small groups will be disregarded certainly. Once a linkage criterion is specified, the clustering method produces as many groups as is required, and the centroid of the largest group gives the information on the location of disturbance without worrying about the aforementioned issue of initial centroids incurred by the K -mean clustering method. Therefore, only hierarchical clustering method will be used in the study, and an agglomerative clustering algorithm would be applied and two types of linkage criteria, average linkage and complete linkage, are used in defining the distance between clusters while following the Euclidean distance metric.

In Fig. 4(a)–(c), we show five clusters obtained in Euclidean space by using the hierarchical clustering analysis for three intrusion events occurring, respectively, at the positions of $d = 1035$, 787 , and 393 m. These positions of intrusion correspond to the position numbers of 0.02 , 4.98 , and 12.86 according to our previous definition of normalized distance from PMPBS2. It should be noted that these five clusters (marked with different colors) were obtained using the complete linkage criterion and that the average linkage criterion would lead to the same five clusters in these cases because of little chain

effect [23] incurred for the two types of linkage criteria in clustering. In each case of intrusion, the centroid of the largest cluster with data points of red triangles is marked by cross with the coordinates shown. Note that the x coordinate of the centroid represents the calculated position number that is essentially obtained by averaging over the data points of the largest cluster. The calculated position number corresponds to a determined location of intrusion, which is obtained here by excluding all uncorrelated data points in Euclidean space. It can then be clearly seen that the discrepancies between the calculated and the real position numbers are 0 , 0.479 , and 0.138 in magnitude for the three cases of intrusion events, which are 0 , 23.95 , and 6.9 m, respectively, in real distance.

C. Effect of Differential Signals on Locating Capability

Lower frequency components of the two detected signals $I_{PD1}(t)$ or $I_{PD2}(t)$ could inherently cause larger errors in calculating the time delay between the two detected signals than higher frequency components. This can be seen from Fig. 2(c) by noting that the calculated position numbers at some frequencies lower than ~ 1000 Hz could reach a magnitude of 20 or even larger, in contrast to those obtained at frequencies higher than 2500 Hz. This was the reason why the position numbers at frequencies higher than 2500 Hz were only considered in determining the location of intrusion in our previous work [19]. Meanwhile, nuisances arising from temperature variation, wind, small animals, and so on could be unwanted sources added to the detected signals and result in locating errors. Such unwanted sources could most likely produce a locating error at the frequencies lower than 1 kHz.

To alleviate the effect of such nuisances, we take a difference operation on every two adjacent signals detected. For example, this operation results in a differential signal $S^i(t) = I_{PD2}^{i+1}(t) - I_{PD2}^i(t)$, where $I_{PD2}^i(t)$ and $I_{PD2}^{i+1}(t)$ are the CW signals $I_{PD2}(t)$ detected in time intervals i and $i + 1$, respectively. Note that $I_{PD2}(t)$ was acquired every 10 ms, and thus, the length of each time interval for $I_{PD2}(t)$ is 10 ms. Similarly, to acquire a differential signal $S^i(t - 2\tau_2)$ from the detected CCW signal $I_{PD1}(t)$, we subtract the CCW signal $I_{PD1}^i(t - 2\tau_2)$ from $I_{PD1}^{i+1}(t - 2\tau_2)$. As noted in Fig. 5, every two adjacent signals coming from CW (or CCW) detection produce a differential signal $S^i(t)$ (or $S^i(t - 2\tau_2)$), where the superscript i refers to time interval i in the sequence of signal acquired. Thus, the FSA method could be applied to determine the time delay $2\tau_2$ now by using the differential signals $S^i(t)$ and $S^i(t - 2\tau_2)$ instead of $I_{PD1}(t)$ and $I_{PD2}(t)$.

We now show the hierarchical clustering results obtained by applying the aforementioned differential signals $S^i(t)$ and $S^i(t - 2\tau_2)$ to calculate the location of intrusion and show the comparison between these results and those based on the use of $I_{PD1}(t)$ and $I_{PD2}(t)$. In doing so, we have consecutively struck the sensing fiber heavily for 1 min at the position of $d = 787$ m (corresponding to the position number 4.98). Fig. 6(a) shows the waveforms of $I_{PD1}(t)$ and $I_{PD2}(t)$ detected for a particular time period of 10 ms, while Fig. 6(b) shows the waveforms detected for the next 10 -ms time period. The waveforms of the differential signals $S(t)$ and $S(t - 2\tau_2)$ are obtained by subtracting the waveforms of $I_{PD1}(t)$ and $I_{PD2}(t)$ in Fig. 6(a)

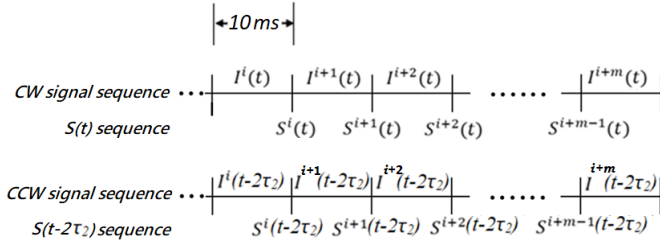


Fig. 5. $S(t)$ is determined by subtracting the CW signal $I^{i+1}(t)$ by $I^i(t)$, where $I^{i+1}(t)$ and $I^i(t)$ are the CW signal $I_{PD2}(t)$ detected in time intervals $i + 1$ and i , respectively. Meanwhile, $S(t - 2\tau_2)$ is determined by subtracting the signal $I^{i+1}(t - 2\tau_2)$ by $I^i(t - 2\tau_2)$, where $I^{i+1}(t - 2\tau_2)$ and $I^i(t - 2\tau_2)$ are the CCW signal $I_{PD1}(t)$ detected in two adjacent time intervals.

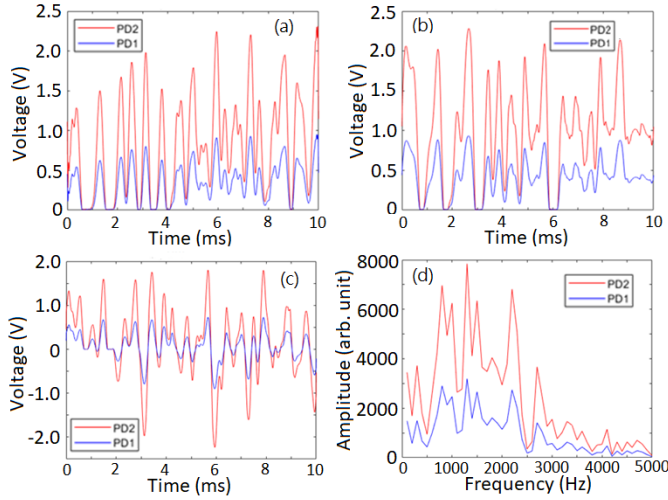


Fig. 6. (a) Signal waveforms detected by PD1 and PD2, i.e., $I_{PD1}(t)$ and $I_{PD2}(t)$, for an intrusion event occurring at $d = 787$ m within a particular time period of 10 ms; (b) signal waveforms detected within the next 10-ms time period; (c) differential signals $S^i(t)$ and $S(t - 2\tau_2)$ obtained by subtracting the two waveforms in (a) from those in (b); and (d) Fourier spectra of $S^i(t)$ and $S(t - 2\tau_2)$.

from those in Fig. 6(b). The two waveforms are denoted by PD2 and PD1 in the legend of Fig. 6(c). The Fourier spectra of $S(t)$ and $S(t - 2\tau_2)$ are shown in Fig. 6(d), denoted by PD2 and PD1, respectively.

The CP value as a function of frequency shown in Fig. 7(a) was obtained from Fig. 6(d). As noted from this figure, the CP values at some frequencies, such as 2500, 3800, 4200, and 5000 Hz, are comparatively small with respect to those at other frequencies. The calculated position number at these frequencies might be unreliable because weak spectral components that have small CP values are susceptible to noises occurring in photodetection and will likely be associated with incorrect position numbers. This can be seen in Fig. 7(b), where the calculated position number versus frequency is shown. As can be seen, the calculated position numbers at, respectively, 2500 and 4200 Hz are ~ 40 and ~ 26 , which are incredibly large (see the spectral components highlighted with circles at these two frequencies). To exclude the position numbers contributed from weak spectral components, we assign a threshold for CP values such that any spectral

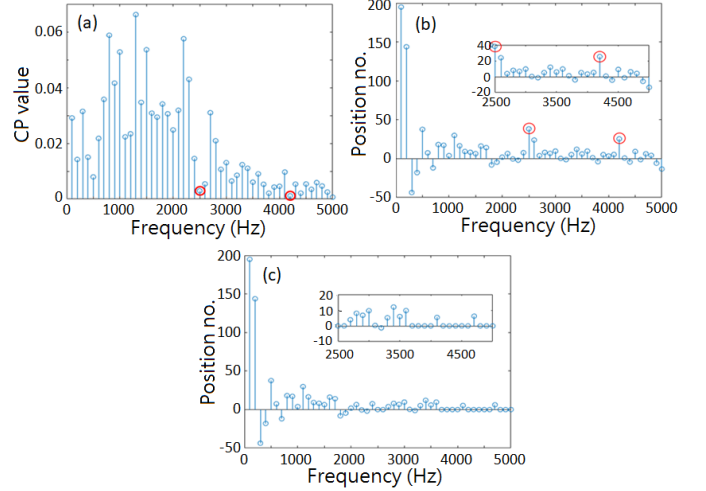


Fig. 7. (a) CP value spectrum obtained from Fig. 6(d), (b) calculated position number as a function of frequency, and (c) spectrum of position number same as (b) except that some position numbers are changed to zero at a certain frequency at which the corresponding CP values are smaller than the threshold $CP_{th} = 0.0055$.

component with a CP value smaller than the threshold will not be considered and neither will the corresponding position number be included for determining the location of intrusion. The threshold value, denoted by CP_{th} hereafter, is set to be the average of all of the CP values in the frequency range of 3100–5000 Hz. The reason for taking this range is that this threshold is mainly used to exclude the weak spectral components beyond a certain frequency, say 3000 Hz. Also, it is noted that the signal detected will mainly fall into the band of 100–3000 Hz, as can be seen from Fig. 6(d). As we know, the determination of the intrusion location is inherently more accurate at the high-frequency band than at the low-frequency band. However, in the high-frequency band, signal amplitudes could be small that detection-noise-induced signal distortion may lead to a locating error. Thus, spectral components with low CP values in the high-frequency band (e.g., from 3100 to 5000 Hz) should not be considered in determining the location of intrusion. In the case of Fig. 6(d), CP_{th} is calculated to be 0.0055. Fig. 7(c) shows the spectrum of position number like Fig. 7(b) except that some position numbers are changed to zero because the corresponding spectral components have CP values smaller than CP_{th} . With these position numbers disregarded, all other position numbers from 100 to 5000 Hz will be considered for subsequent clustering analysis.

Fig. 8(a) and (b) shows five clusters (appearing in different colors) obtained in the Euclidean space by using the hierarchical clustering analysis with average linkage criterion employed for an intrusion event occurring at the position of $d = 787$ m (corresponding to the position number of 4.98). They are the results derived from Fig. 6(c). Fig. 8(a) shows the result by considering all of the 50 spectral components from 100 to 5000 Hz, while Fig. 8(b) shows the result by excluding the spectral components with CP values lower than CP_{th} in the same frequency range. The largest cluster in either case contains data points of red triangles. The x coordinates of the two centroids marked by crosses in Fig. 8(a) and (b) read 5.208 and 5.165, respectively, which represent the determined

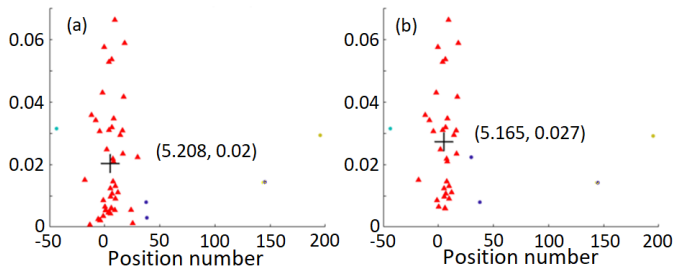


Fig. 8. Five clusters (appearing in different colors) obtained in Euclidean space by using the hierarchical clustering analysis with average linkage criterion employed for an intrusion event occurring at the position of $d = 787$ m, i.e., for the case of Fig. 7. (a) Result with all 50 data points in the spectrum of CP value considered. (b) Result by excluding the data points with lower CP values. These results are obtained by using the differential signals in Fig. 6(c). The largest cluster in either case contains data points of red triangles.

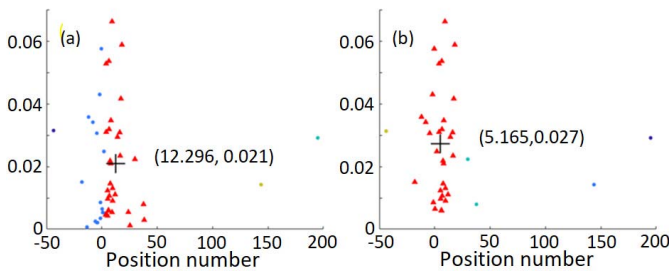


Fig. 9. Five clusters (appearing in different colors) obtained in Euclidean space by using the hierarchical clustering analysis with complete linkage criterion employed for the case of Fig. 7. (a) Result with all 50 data points in the spectrum of CP value considered. (b) Result by excluding the data points with lower CP values. These results are obtained by using the differential signals in Fig. 6(c). The largest cluster in either case contains data points of red triangles.

locations of intrusion in the two cases. Notably, the locating error is a little bit smaller when the spectral components with CP values smaller than CP_{th} are excluded, compared to the case without excluding any spectral components in the spectral range of 100–5000 Hz. For the same case of Fig. 6(c), the hierarchical clustering analysis shows a quite different result as complete linkage criterion is employed. Fig. 9(a) shows the result with all 50 spectral components from 100 to 5000 Hz considered, while Fig. 9(b) shows the result by excluding the spectral components with lower CP values. It is obvious that quite an erroneous location of intrusion is determined when all 50 data points are considered in the Euclidean space, while the determined position number is the same as that obtained by using the average linkage criterion for the case of excluding the weak spectral components with CP values smaller than CP_{th} .

Note that the x coordinate of the centroid of the largest cluster represents a mean position number by counting all of the data points in the largest cluster. It should also be noted that these data points wander around the centroid, meaning that the associated position numbers fluctuate around their mean. For example, the data points in the largest cluster shown in Fig. 8(b) have their x coordinates fluctuating in the range of -18.12 to 18.25 around the mean of 5.165 .

For comparison, we show below the clustering results obtained by using the two signal waveform pairs $[I^i(t), I^i(t - 2\tau_2)]$ and $[I^{i+1}(t), I^{i+1}(t - 2\tau_2)]$, which are the

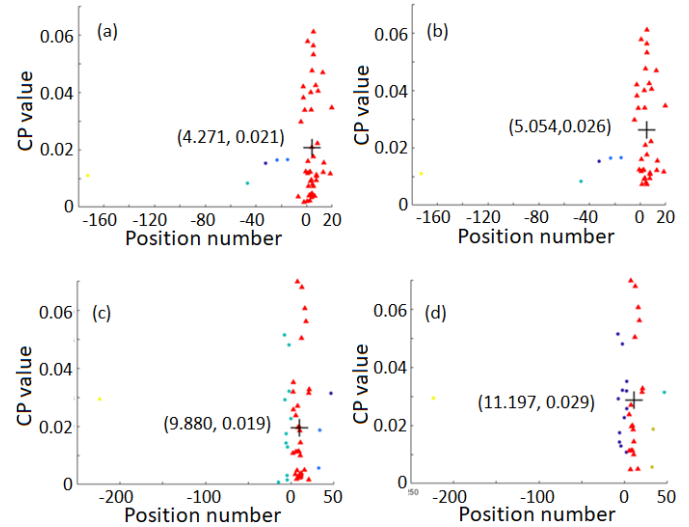


Fig. 10. Five clusters (appearing in different colors) obtained in Euclidean space by using the hierarchical clustering analysis with average linkage criterion employed. (a) Five clusters with all 50 data points in the spectrum of CP value considered. (b) Five clusters formed by excluding data points with CP values lower than $CP_{th} = 0.0065$, for the case of Fig. 6(a). (c) Clustering result with all 50 data points in the spectrum of CP value considered. (d) Five clusters formed by excluding data points with CP values lower than $CP_{th} = 0.0042$, for the case of Fig. 6(b). The largest cluster in each case contains data points of red triangles.

signal pair $[I_{PD1}(t), I_{PD2}(t)]$ detected within a particular time interval and its subsequent time interval, as shown in Fig. 5. With average linkage criterion employed, the clustering result for the signal waveforms shown in Fig. 6(a) in the case of including all 50 data points is shown in Fig. 10(a). On the other hand, Fig. 10(b) shows the clustering result when the spectral components with CP values smaller than $CP_{th} = 0.0065$ are excluded. As can be seen, the x coordinates of the centroids of the largest clusters in the two cases read 4.271 and 5.054 . This indicates that the locating error derived by using the signal waveform pair $[I^i(t), I^i(t - 2\tau_2)]$ with all weak spectral components excluded is smaller than the case with all 50 spectral components taken into consideration. However, the locating error for the case of using the signal waveform pair $[I^{i+1}(t), I^{i+1}(t - 2\tau_2)]$, i.e., the signal waveforms in Fig. 6(b), is incredibly large with or without excluding all weak spectral components. This can be seen from Fig. 10(c) and (d), where the determined position numbers read 9.88 and 11.197 for the cases without and with weak spectral components excluded, respectively. The locating errors are 4.9 and 6.217 in position number, corresponding to 245 and 310.85 m in real distance, respectively.

When a complete linkage criterion is used for the same signal waveforms, the clustering analysis for obtaining five clusters could show a different result, which are likewise inaccurate though. Fig. 11(a) shows five clusters (appearing in different colors) with all 50 data points in the spectrum of CP value considered, while Fig. 11(b) shows five clusters formed by excluding data points with CP values lower than $CP_{th} = 0.0065$ for the case of the signal waveforms in Fig. 6(a). Obviously, the determined locations of intrusion read 2.875 and 2.070 in terms of position number, and these

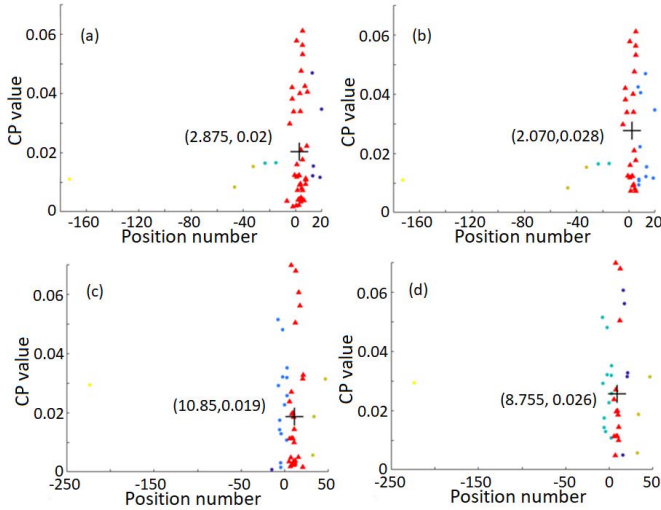


Fig. 11. Five clusters (appearing in different colors) obtained in Euclidean space by using the hierarchical clustering analysis with complete linkage criterion. (a) Five clusters with all 50 data points in the spectrum of CP value considered. (b) Five clusters formed by excluding data points with CP values lower than $CP_{th} = 0.0065$, for the case of Fig. 6(a). (c) Clustering result with all 50 data points in the spectrum of CP value considered. (d) Five clusters formed by excluding data points with CP values lower than $CP_{th} = 0.0042$, for the case of Fig. 6(b). The largest cluster in each case contains data points of red triangles.

represent a big deviation from the real value of 4.98. On the other hand, Fig. 11(c) and (d) shows the clustering results without and with, respectively, weak spectral components excluded, for the case of Fig. 6(b). Again, one can see that the determined position numbers, which are represented by the x coordinates of the centroids of the largest clusters in both cases, are far from being accurate.

From the results above, the average linkage criterion could sometimes be better than the complete linkage criterion, but in some cases turns to be worse in locating the intrusion event for the five cluster analysis, as indicated by the results in Figs. 10 and 11. The clustering analysis fails sometimes even when the weak spectral components are excluded, as can be seen from Figs. 10(d) and 11(b) and (d). On the other hand, Figs. 8 and 9 show that both average linkage and complete linkage criteria could improve the locating accuracy when differential signals are used for clustering analysis. Use of either criterion leads to a locating error of 0.185, which is equivalent to 9.25 m in real distance. Thus, when using differential signals for locating an intrusion event, one can obtain more accurate location of intrusion than what is obtained by using directly detected signals.

To show that differential signals also work for intrusion events occurring at other locations, we show five clusters formed in Euclidean space by using the hierarchical clustering analysis with average linkage criterion applied to a given pair of differential signals $S(t)$ and $S(t - 2\tau_2)$, for intrusion events occurring at the positions of $d = 1035$ m (corresponding to the position number 0.02) and $d = 393$ m (corresponding to the position number 12.86). Fig. 12(a) shows five clusters with all 50 spectral components considered, while Fig. 12(b) shows five clusters obtained by excluding weak spectral components

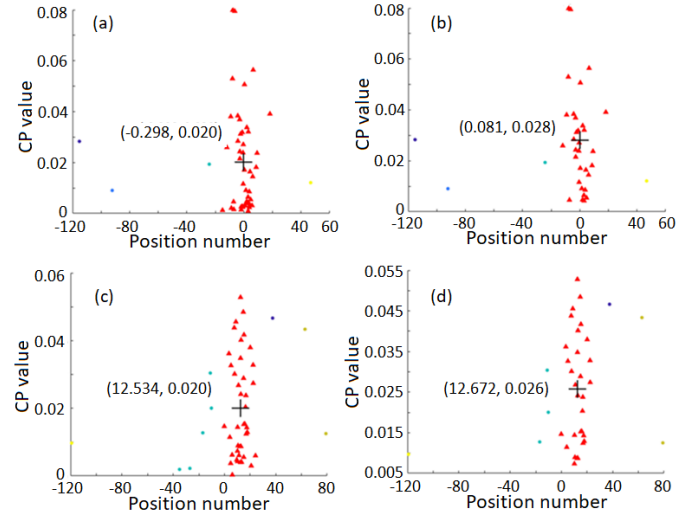


Fig. 12. Five clusters (appearing in different colors) obtained in Euclidean space by using the hierarchical clustering analysis with average linkage criterion applied to a particular pair of differential signals $S(t)$ and $S(t - 2\tau_2)$. (a) Five clusters with all 50 spectral components considered. (b) Five clusters obtained by excluding weak spectral components with CP values lower than $CP_{th} = 0.0041$, for an intrusion event occurring at $d = 1035$ m. (c) Clustering result with all 50 spectral components considered. (d) Five clusters obtained by excluding weak spectral components with CP values lower than $CP_{th} = 0.0069$, for an intrusion event occurring at $d = 393$ m. The largest cluster in each case contains data points of red triangles.

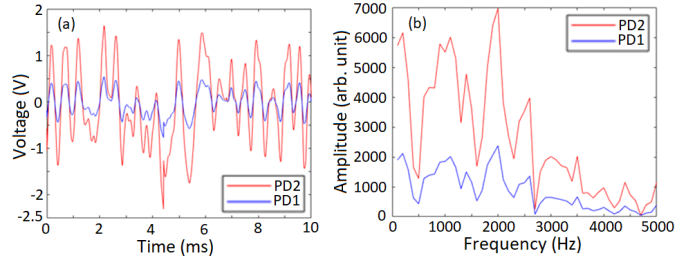


Fig. 13. (a) Signal waveforms of differential signals $S(t)$ and $S(t - 2\tau_2)$. (b) Fourier spectra of $S(t)$ and $S(t - 2\tau_2)$. The case is for $d = 393$ m.

that have CP values lower than $CP_{th} = 0.0041$, for an intrusion event occurring at $d = 1035$ m. Fig. 12(c) shows the clustering result with all 50 spectral components considered, while Fig. 12(d) shows five clusters obtained by excluding weak spectral components that have CP values lower than $CP_{th} = 0.0069$, for an intrusion event occurring at $d = 393$ m. From the x coordinate of the centroid of the largest cluster, one can read the determined position numbers 0.081 and 12.672 for the cases of excluding contributions from weak spectral components, in contrast to -0.298 and 12.534 for the cases without excluding weak spectral components. When the weak spectral components are excluded, the locating errors for the cases of Fig. 12(b) and (d) are no larger than 3.05 and 9.4 m, respectively, in real distance.

D. Effect of Number of Clusters

If the number of clusters is small, some uncorrelated data points in Euclidean space may group together, resulting in a big cluster. These data points may represent either some

TABLE I

DETERMINED POSITION NUMBER AND LOCATING ERROR FOR AN INTRUSION EVENT OCCURRING AT THE POSITION OF $d = 393$ M FOR VARIOUS N_c 'S. HERE, THE CALCULATION IS BASED ON THE USE OF THE SIGNAL WAVEFORMS SHOWN IN FIG. 13(A) AND THE AVERAGE LINKAGE CRITERION

No. of clusters	X coordinate of the centroid of the largest cluster	Locating error in position number
3	11.162 (9.601)*	-1.698 (-3.259)
4	13.474 (13.129)	0.614 (0.269)
5	12.672 (12.535)	-0.188 (-0.326)
6	12.672 (12.534)	-0.188 (-0.326)
7	14.140 (12.534)	1.28 (-0.325)
8	13.165 (11.224)	0.305 (-1.636)

* The numbers in parentheses denote the results obtained without excluding any spectral component.

weak spectral components with small CP values or strong spectral components in a low-frequency band, such as that from 100 to 2500 Hz. The data points corresponding to these two kinds of spectral components may lead to a large shift in the position of the centroid of interest and, accordingly, a wrong position of intrusion event. Here, we will compare the clustering results obtained by choosing 3–8 for the number of clusters (denoted by N_c hereafter) for an intrusion event occurring at the position of $d = 393$ m (corresponding to the position number 12.86). Note that the differential signals for such an intrusion event have been used for demonstrating clusters in Fig. 12(c) and (d) for $N_c = 5$. Fig. 13(a) shows the differential signals of interest, with their Fourier amplitude spectra shown in Fig. 13(b). Using this pair of differential signals and excluding all weak spectral components that have CP values smaller than $CP_{th} = 0.0069$, 3–8 clusters were obtained in the Euclidean space with average linkage criterion employed. In each case of N_c ($N_c = 3$ –8), the x coordinate of the centroid of the largest cluster, which corresponds to the determined location of intrusion, was found and listed in Table I. Clearly, a value of N_c equal to 5 or 6 would lead to a minimum location error of 0.188 in magnitude expressed in position number or 9.4 m in real distance. For comparison, Table I also shows in parentheses the results obtained without excluding any spectral component. For the latter case, the minimum locating error 0.269 (or 13.45 m in real distance) was obtained at $N_c = 4$, which was a little bit larger than 0.188 obtained for the case with weak spectral components excluded.

Using complete linkage criterion for the same intrusion event, we obtain the locations of intrusion for various N_c 's, as shown in Table II. The minimum locating error occurs at $N_c = 5$, which gives 9.4 m for the locating error. If we choose any other N_c value, we can see that the locating accuracy would not be better than that obtained with average linkage criterion employed for this intrusion event. The numbers enclosed in parentheses represent the results obtained without excluding any spectral component. In this example, when N_c is chosen to be 3 to 7, all of the locating errors would be larger than that at $N_c = 5$ in the case with weak spectral components excluded. However, at $N_c = 8$, the locating error reaches 0.059 for this particular case. A comparison between

TABLE II

DETERMINED POSITION NUMBER AND LOCATING ERROR FOR AN INTRUSION EVENT OCCURRING AT THE POSITION OF $d = 393$ M FOR VARIOUS N_c 'S. HERE, THE CALCULATION IS BASED ON THE USE OF SIGNAL WAVEFORMS SHOWN IN FIG. 13(A) AND THE COMPLETE LINKAGE CRITERION

No. of clusters	X coordinate of the centroid of the largest cluster	Locating error in position number
3	11.162 (17.941)*	-1.698 (5.081)
4	13.474 (14.915)	0.614 (2.055)
5	12.672 (14.915)	-0.188 (2.055)
6	9.087 (14.250)	-3.773 (1.39)
7	9.087 (14.250)	-3.773 (1.39)
8	10.918 (12.919)	-1.942 (0.059)

* The numbers in parentheses denote the results obtained without excluding any spectral component.

the results in Tables I and II reveal that the average linkage criterion would produce a better or the same locating accuracy with respect to the complete linkage criterion. This comparison again demonstrates that weak spectral components should be excluded in determining the location of intrusion. Furthermore, the minimum locating error can be reached by choosing N_c as 5 or 6 for the average linkage criterion for an intrusion event occurring at the position of $d = 393$ m when weak spectral components are excluded. For intrusion events occurring at other positions, the same conclusion can almost apply. For example, for an intrusion event occurring at the position of $d = 1035$ m (corresponding to the position number 0.02), a minimum locating error of 0.061 (or 3.05 m in real distance) can be reached at $N_c = 4$ or 5 in contrast to the second smallest locating error of 0.523 (or 26.15 m) at $N_c = 6$. The locating errors at $N = 3, 7$, and 8 are 0.677 (or 33.85 m), 1.388 (or 69.4 m), and 4.373 (or 218.65 m), respectively, for this event. On the other hand, for an intrusion event occurring at the position of $d = 787$ m (corresponding to the position number 4.98), a minimum locating error of 0.185 in position number occurs at $N_c = 5$ in contrast to the second smallest locating error of 0.194 at $N = 7$ or 8. The locating errors at $N = 3, 4$, and 6 are 0.431, 1.917, and 2.528, respectively, for this event. Therefore, we will choose $N = 5$ for subsequent discussion. Both the average linkage criterion and the complete linkage criterion will be compared in terms of locating accuracy.

III. AVERAGE METHOD

A. Locating an Intrusion Event by Averaging Method

It should be noted that all results obtained previously were based on the analysis of a particular intrusion event occurring within a time period of 10 ms. However, an intrusion event should not be determined by merely examining the signal waveforms acquired within a particular 10-ms time period or several 10-ms time periods. In a previous work [19], we have investigated the locating capability of the DMZI intrusion sensor by examining 100 pairs of signal waveforms (i.e., CW and CCW signal waveforms) with each pair acquired in a 10-ms time period, where an intrusion event was defined if more than a half of the 100 pairs of signal waveforms were reduced to an intrusion case in which three parameters (i.e., SA, LC, and FR [19]) all exceeded their respective thresholds.

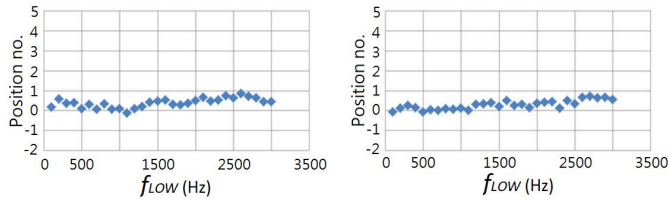


Fig. 14. Determined position number as a function of f_{LOW} obtained by using complete linkage criterion (left) and average linkage criterion (right) in the clustering analysis with $N_c = 5$, for an intrusion event occurring at the position of $d = 1035$ m (corresponding to the position number 0.02). Each data point is obtained by averaging the 99 position numbers with each obtained by applying the clustering analysis to a pair of differential signals $S(t)$ and $S(t - 2\tau_2)$. For each data point, spectral components in a frequency range of f_{LOW} to 5000 Hz were used, however, with weak spectral components with CP values smaller than CP_{th} excluded.

By averaging the calculated locations of intrusion over the majority of intrusion cases, we then obtained the location of intrusion for this intrusion event that lasts for 1 s. In this study, 100 pairs of $IP_{D1}(t)$ and $IP_{D2}(t)$ are used to produce 99 pairs of $S(t)$ and $S(t - 2\tau_2)$. We then apply the clustering analysis to each pair of $S(t)$ and $S(t - 2\tau_2)$ and obtain 99 clustering results with each x coordinate of the centroid of the largest cluster representing the determined location of intrusion for the pair of concern. We then obtain 99 determined locations of intrusion. Note that a minority of the determined locations of intrusion may be inaccurate for an intrusion event of concern. However, most determined locations of intrusion are accurate for the intrusion event. Then, by averaging all of the 99 determined locations of intrusion, we obtained the location of intrusion for such an intrusion event occurring in a time period of 1 s. In this study, we continuously knock the fiber cable at a given position for a second to simulate an intrusion event.

It is noted that the hierarchical clustering analysis was performed after weak spectral components in the range of 100–5000 Hz were excluded in some previous examples. We are then curious to see whether the clustering analysis provides a better estimation for the location of intrusion if a different frequency range is used with weak spectral components being excluded. Since lower frequency spectral components may provide larger locating errors than higher frequency ones, as revealed in Fig. 7(b) or (c). Then, locating errors might be reduced if some lower frequency spectral components are not chosen for consideration in the first place. For example, taking the spectral components in a range of 500–5000 Hz for clustering analysis could be better than taking the whole range, i.e., from 100 to 5000 Hz. Here, we discuss this issue by considering a variety of spectral ranges for the spectral components in the clustering analysis. Each spectral range starts from f_{LOW} to 5000 Hz in the clustering analysis, where f_{LOW} varies from 100 to 3000 Hz with an interval of 100 Hz.

Fig. 14 shows the determined position number as a function of f_{LOW} obtained by using the complete linkage criterion (left) and average linkage criterion (right) in the clustering analysis with $N_c = 5$, for an intrusion event occurring at the position of $d = 1035$ m (corresponding to the position number 0.02). Here, each data point was obtained by averaging 99 calculated position numbers, each of which was obtained by applying

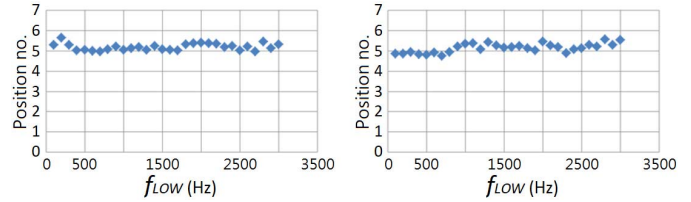


Fig. 15. Determined position number as a function of f_{LOW} obtained by using complete linkage criterion (left) and average linkage criterion (right) in the clustering analysis with $N_c = 5$, for an intrusion event occurring at the position of $d = 787$ m (corresponding to the position number 4.98). Each data point is obtained by averaging the 99 position numbers with each obtained by applying the clustering analysis to a pair of differential signals $S(t)$ and $S(t - 2\tau_2)$. For each data point, spectral components in a frequency range of f_{LOW} to 5000 Hz were used, however, with weak spectral components with CP values smaller than CP_{th} excluded.

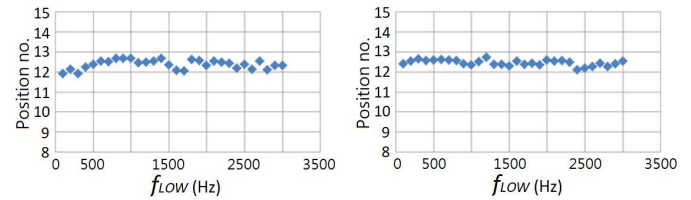


Fig. 16. Determined position number as a function of f_{LOW} obtained by using complete linkage criterion (left) and average linkage criterion (right) in the clustering analysis with $N_c = 5$, for an intrusion event occurring at the position of $d = 393$ m (corresponding to the position number 12.86). Each data point is obtained by averaging the 99 position numbers with each obtained by applying the clustering analysis to a pair of differential signals $S(t)$ and $S(t - 2\tau_2)$. For each data point, spectral components in a frequency range of f_{LOW} to 5000 Hz were used, however, with weak spectral components with CP values smaller than CP_{th} excluded.

the clustering analysis to a pair of differential signals $S(t)$ and $S(t - 2\tau_2)$. It can be seen that high locating accuracy could be reached if a proper f_{LOW} is chosen for both linkage criteria.

To see how these two linkage criteria work for intrusion events occurring at different positions, we show the determined position numbers in Figs. 15 and 16 for intrusion events occurring at the positions of $d = 787$ m (corresponding to the position number 4.98) and $d = 393$ m (corresponding to the position number 12.86), respectively. In both figures, the determined position numbers as a function of f_{LOW} with complete linkage criterion are shown on the left, while those obtained by using average linkage criterion are shown on the right. By using the averaging method with a proper f_{LOW} chosen, we can obtain quite a good locating accuracy. For example, when the average linkage criterion is employed, a locating accuracy to be within 0.235 (i.e., 11.75 m in real distance; see Fig. 16 for this number), for the three aforementioned intrusion events, can be reached for $f_{LOW} = 500$ or 600 Hz, in contrast to the case of $f_{LOW} = 100$ Hz with an accuracy to be within 0.41 (i.e., 20.5 m in real distance; see Fig. 16 for this number). On the other hand, when a complete linkage criterion is employed, f_{LOW} can be chosen to be 1000 Hz to have locating errors less than 0.07 (i.e., 3.5 m in real distance; see Fig. 15) for the three intrusion events. For comparison, the locating error can reach 0.751 (i.e., 37.55 m) for an intrusion event occurring at the position of $d = 787$ m for $f_{LOW} = 200$ Hz (see Fig. 15) with the same linkage criterion used.

TABLE III
MAE OBTAINED FOR NUMBER OF CLUSTERS (DENOTED BY N_c) VARYING FROM 3 TO 8 WHEN AN INTRUSION EVENT OCCURS AT THE POSITIONS OF 0.02, 4.98, AND 12.86 WITH COMPLETE LINKAGE CRITERION USED

Location of intrusion (in position number)	MAE obtained when N_c is equal to					
	3	4	5	6	7	8
0.02	0.258	0.172	0.212	0.310	0.363	0.277
4.98	0.184	0.222	0.193	0.158	0.198	0.218
12.86	0.153	0.301	0.270	0.308	0.399	0.276
	(1100)*	(1000)*	(1200)*	(1600)*	(800)*	(1800)*

* The numbers in parentheses represent the lower bound frequency f_{LOW} that is used for calculation of each MAE for a given N_c .

B. Reliability Test

Both linkage criteria have proved to be able to determine the location of intrusion with an accuracy to be within 11.75 m (for average linkage criterion) or 3.5 m (for complete linkage criterion) when a proper f_{LOW} was chosen as the lower bound of the spectral range for the Fourier components of the signal. In the following, we will check the reliability of the clustering analysis method by examining the locating capability when an intrusion event occurs at a given position. Still, the test is carried out for three positions, i.e., the positions of 0.02, 4.98, and 12.86. In simulating an intrusion event at a given position, we knock the sensing fiber continuously for 1 s. At each position, we then repeat this knocking to produce 12 intrusion events totally. After the 12 intrusion events are analyzed, we can obtain a set of 12 determined locations of intrusion for each position, from which we can then obtain a mean absolute error (MAE) defined in the following equation for the determined location of intrusion:

$$MAE(j) = \frac{1}{12} \cdot \sum_{i=1}^{12} |\varphi_{ji} - \bar{\varphi}_j| \quad (5)$$

where φ_{ji} is the i th determined location of intrusion ($i = 1, 2, \dots, 12$) for an intrusion event occurring at the j th location ($j = 1, 2, 3$) and $\bar{\varphi}_j$ is the true j th location of intrusion.

Table III shows the MAEs obtained for N_c (number of clusters) varying from 3 to 8 when an intrusion event occurs at three of the aforementioned positions with complete linkage criterion used in the clustering analysis, where the numbers in parentheses represent the lower bound frequency f_{LOW} that is used for calculation of each MAE for a given N_c . It should be stated that each f_{LOW} shown here gives the minimum average value of MAEs for a given N_c . It appears that MAEs could be no larger than 0.258 for the three cases of intrusion locations with $N_c = 3$ if spectral components from 1100 to 5000 Hz are considered only, and this MAE corresponds to an average error of 12.9 m in real distance. On the other hand, Table IV shows the MAEs when an intrusion event occurs at the positions of 0.02, 4.98, and 12.86 with average linkage criterion used in the clustering analysis. Again, each f_{LOW} shown in the parentheses gives the minimum average value of MAEs for the three cases of intrusion locations for a given N_c . The MAEs achieved for each N_c are a little smaller than those shown

TABLE IV
MAE OBTAINED FOR A NUMBER OF CLUSTERS (DENOTED BY N_c) VARYING FROM 3 TO 8 WHEN AN INTRUSION EVENT OCCURS AT THE POSITIONS OF 0.02, 4.98, AND 12.86 WITH COMPLETE LINKAGE CRITERION USED. THE SIGNALS $I_{PD1}(t)$ AND $I_{PD2}(t)$ APPLY TO THE CLUSTERING ANALYSIS

Location of intrusion (in position number)	MAE obtained when N_c is equal to					
	3	4	5	6	7	8
0.02	0.201	0.196	0.164	0.237	0.138	0.176
4.98	0.204	0.256	0.231	0.240	0.273	0.288
12.86	0.256	0.216	0.177	0.227	0.286	0.291
	(1100)*	(500)*	(500)*	(500)*	(200)*	(100)*

* The numbers in parentheses represent the lower bound frequency f_{LOW} that is used for calculation of each MAE for a given N_c .

TABLE V
MAE OBTAINED FOR NUMBER OF CLUSTERS (DENOTED BY N_c) VARYING FROM 3 TO 8 WHEN AN INTRUSION EVENT OCCURS, RESPECTIVELY, AT THE POSITIONS OF 0.02, 4.98, AND 12.86 WITH COMPLETE LINKAGE CRITERION USED. THE SIGNALS $I_{PD1}(T)$ AND $I_{PD2}(T)$ APPLY TO THE CLUSTERING ANALYSIS

Location of intrusion (in position number)	MAE obtained when N_c is equal to					
	3	4	5	6	7	8
0.02	0.250	0.283	0.252	0.384	0.368	0.204
4.98	0.242	0.279	0.329	0.267	0.402	0.211
12.86	0.302	0.425	0.349	0.369	0.410	0.422
	(900)*	(1000)*	(300)*	(300)*	(400)*	(400)*

* The numbers in parentheses represent the lower bound frequency f_{LOW} that is used for calculation of each MAE for a given N_c .

TABLE VI
MAE OBTAINED FOR NUMBER OF CLUSTERS (DENOTED BY N_c) VARYING FROM 3 TO 8 WHEN AN INTRUSION EVENT OCCURS, RESPECTIVELY, AT THE POSITIONS OF 0.02, 4.98, AND 12.86 WITH AVERAGE LINKAGE CRITERION USED. THE SIGNALS $I_{PD1}(T)$ AND $I_{PD2}(T)$ APPLY TO THE CLUSTERING ANALYSIS

Location of intrusion (in position number)	MAE obtained when N_c is equal to					
	3	4	5	6	7	8
0.02	0.268	0.262	0.261	0.415	0.460	0.352
4.98	0.293	0.258	0.317	0.336	0.382	0.341
12.86	0.418	0.499	0.285	0.479	0.375	0.447
	(1200)*	(700)*	(600)*	(500)*	(1200)*	(500)*

* The numbers in parentheses represent the lower bound frequency f_{LOW} that is used for calculation of each MAE for a given N_c .

in the case of Table III, where a complete linkage criterion is used. The best N_c that could lead to minimum values of MAEs is 5, and these MAEs are no larger than 0.231. Also, this means that the mean error for the total 36 intrusion events occurring at the three intrusion locations could be all smaller than 11.55 m if N_c is chosen at 5, the average linkage criterion is employed, and $f_{LOW} = 500$ Hz is set for use.

For comparison, we have also estimated the MAEs based on the use of the signals $I_{PD1}(t)$ and $I_{PD2}(t)$ in the clustering analysis. Tables V and VI show the MAEs obtained by using the complete linkage criterion and average linkage criterion, respectively, for intrusion events occurring at three positions. The best N_c is 3 (or 5) when the complete linkage criterion (or average linkage criterion) is used in the clustering analysis

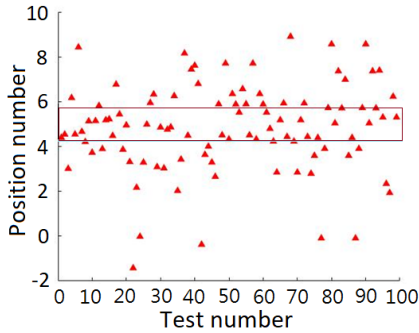


Fig. 17. Distribution of the 99 position numbers calculated by hierarchical clustering method ($N_c = 5$) with average linkage criterion used for each of the 99 pairs of differential signals, in the case of an intrusion occurring at the position of 4.98. The rectangular box encircles the position numbers near the mean 4.814 within the accuracy of ± 0.75 .

with the signals $I_{PD1}(t)$ and $I_{PD2}(t)$ applied. The MAE is no larger than 0.302 (or 0.317) for $N_c = 3$ (or 5) when the complete linkage criterion (or average linkage criterion) is used for locating the intrusion. Both criteria lead to a little larger MAE compared with that achieved by using the average linkage criterion for $N_c = 5$ when the differential signals $S(t)$ and $S(t - 2\tau_2)$ are applied (see Table IV).

It should be noted that some of the 100 pairs of signal waveforms acquired in the duration of 1 s do not correspond to an intrusion case, and their determined locations of intrusion can be inaccurate. Fig. 17 shows the distribution of the 99 position numbers calculated by the hierarchical clustering method ($N_c = 5$) with average linkage criterion used for each of the 99 pairs of differential signals, in the case of an intrusion occurring at the position of 4.98. The average of the 99 calculated position numbers is 4.814 for this intrusion event, corresponding to a locating error of 0.166. The rectangular box in Fig. 17 encircles 40 position numbers that are near 4.814 within the accuracy of ± 0.75 . Many calculated position numbers residing outside the box correspond to a larger locating error than those inside the box. Also, almost all of them result from the signal waveforms that represent nonintrusion cases. The reason for the existence of these nonintrusion signals can be explained as follows. When the fiber cable was continuously knocked for 1 s to simulate an intrusion event, the type of knocking was intermittent, and there existed a number of signal waveforms during the time interval when the fiber cable was not actually knocked or just slightly vibrated. In the practical situation of climbing a netted fence, the fiber cable is vibrated intermittently as well.

IV. CONCLUSION

The hierarchical clustering analysis is presented for the first time here to locate the intrusion-induced disturbance on the sensing fibers of a DMZI system. Two linkage criteria, i.e., complete linkage criterion and average linkage criterion, were used in the clustering analysis. It was found that the average linkage criterion provided only a minor improvement in locating capability in terms of MAE with respect to the complete linkage criterion, as can be seen from Table III (see the case of $N_c = 3$) and Table IV (see the case of $N_c = 5$). Based on the evaluation of MAE, we also found that the use of

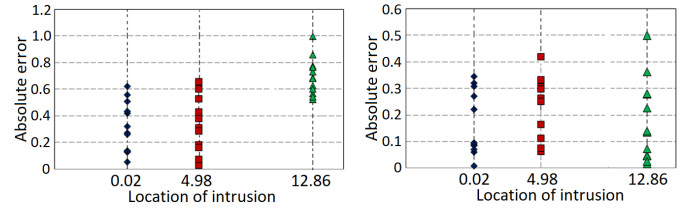


Fig. 18. Absolute locating errors for 12 intrusion events occurring at each of the three positions 0.02, 4.98, and 12.86. The absolute errors obtained by using our previous method are shown in the left, while those obtained by using the presented clustering analysis method are shown in the right.

differential signals for clustering analysis could provide better locating capability than the case when the directly detected signals $I_{PD1}(t)$ and $I_{PD2}(t)$ were used, no matter which of the two linkage criteria was used, as can be seen by comparing Fig. 3 (or Fig. 4) with Fig. 5 (or Fig. 6). By examining the clustering results shown in Figs. 8, 9, and 12, we also found that the locating accuracy was improved for the case of excluding weak spectral components with respect to the case when weak spectral components were not excluded, when differential signals were used for the clustering analysis.

As mentioned previously, we have used 100 pairs of signals to address every intrusion event lasting for 1 s. We could use fewer or more pairs to define how long an intrusion lasted, without any difficulty in dealing with mathematical computation. However, as noted from the work of [19], when an intruder's action is applied on a netted fence, the time consumed to complete an action, such as climbing, breaking, or cutting, is at least 1 s. Also, when a stone or a ball is maliciously thrown at the fence, the fiber cable would be heavily vibrated for at least half a second, that is, the determination of an intrusion and the locating of the intrusion can be done by using 100 pairs of signal waveforms. Also, besides, the use of a larger number of pairs of signal waveforms infers a need of taking longer to wait for an alarm to go off once an intrusion is detected, which is not usually expected.

Note that we have previously presented an estimation method for the location of intrusion for the DMZI sensing system [19], where a test of five intrusion events at each of the two locations 0.02 and 4.98 was conducted; 100 pairs of detected signals $I_{PD1}(t)$ and $I_{PD2}(t)$ for each intrusion event were used to determine the location of intrusion, and a locating error of 0.53 (or 0.3078) was found for an intrusion at the position of 0.02 (or 4.98). The previous locating method is obviously worse than that presented in this study. To further look at this subject, we have followed our previous method and calculated the absolute values of locating error for the same 36 intrusion events with 12 events occurring at each of the three positions. These absolute errors are shown in the left of Fig. 18. We can see that the absolute errors for the 12 intrusion events occurring at the positions of 0.02 and 4.98 range from 0.025 to 0.651, while the absolute errors are between 0.516 and 0.985 for intrusions at the position of 12.86. These absolute errors give MAEs of 0.321, 0.317, and 0.695 for intrusions at the positions of 0.02, 4.98, and 12.86, respectively. In comparison, we show the 12 absolute errors obtained by using the average linkage criterion with

$f_{\text{LOW}} = 500$ Hz and $N_c = 5$ for each position in the right of Fig. 18, where we can see that the 12 absolute errors for the intrusion at the position of 12.86 spread over the widest range, i.e., a range from 0.011 to 0.494, which leads to an MAE of 0.177. The other MAEs are 0.164 and 0.231 for intrusions at the positions of 0.02 and 4.98, respectively. These MAEs are all smaller than those obtained by using our previous method. In conclusion, we have presented a clustering algorithm to achieve a more accurate determination of intrusion location.

REFERENCES

- [1] Y. Lu, T. Zhu, L. Chen, and X. Bao, "Distributed vibration sensor based on coherent detection of phase-OTDR," *J. Lightw. Technol.*, vol. 28, no. 22, pp. 3243–3249, Nov. 15, 2010.
- [2] A. Masoudi, M. Belal, and T. P. Newson, "A distributed optical fibre dynamic strain sensor based on phase-OTDR," *Meas. Sci. Technol.*, vol. 24, no. 8, Jul. 2013, Art. no. 085204.
- [3] F. Peng, H. Wu, X.-H. Jia, Y.-J. Rao, Z.-N. Wang, and Z.-P. Peng, "Ultra-long high-sensitivity ϕ -OTDR for high spatial resolution intrusion detection of pipelines," *Opt. Exp.*, vol. 22, no. 11, pp. 13804–13810, May 2014.
- [4] M. Aktas, T. Akgun, M. U. Demircin, and D. Buyukaydin, "Deep learning based multi-threat classification for phase-OTDR fiber optic distributed acoustic sensing applications," *Proc. SPIE*, vol. 10208, Apr. 2017, Art. no. 102080G.
- [5] N. Yang, Y. Zhao, and J. Chen, "Real-time ϕ -OTDR vibration event recognition based on image target detection," *Sensors*, vol. 22, no. 3, pp. 1127–1149, Feb. 2022.
- [6] P. R. Hoffman and M. G. Kuzyk, "Position determination of an acoustic burst along a Sagnac interferometer," *J. Lightw. Technol.*, vol. 22, no. 2, pp. 494–498, Feb. 2004.
- [7] W. Xu, C. Zhang, S. Liang, L. Li, W. Lin, and Y. Yang, "Fiber-optic distributed sensor based on a Sagnac interferometer with a time delay loop for detecting time-varying disturbance," *Microw. Opt. Technol. Lett.*, vol. 51, no. 11, pp. 2564–2567, Nov. 2009.
- [8] X. Hong, J. Wu, C. Zuo, F. Liu, H. Guo, and K. Xu, "Dual Michelson interferometers for distributed vibration detection," *Appl. Opt.*, vol. 50, no. 22, pp. 4333–4338, Aug. 2011.
- [9] Q. Li, H. Wang, L. Li, S. Liang, and X. Zhong, "Fiber-optic sensor based on Michelson interferometers for distributed disturbance detection," *Infr. Laser Eng.*, vol. 44, no. 1, pp. 205–209, Jan. 2015.
- [10] S. J. Spammer, P. L. Swart, and A. A. Chtcherbakov, "Merged Sagnac-Michelson interferometer for distributed disturbance detection," *J. Lightw. Technol.*, vol. 15, no. 6, pp. 972–976, Jun. 1997.
- [11] Q. Song et al., "Improved localization algorithm for distributed fiber-optic sensor based on merged Michelson-Sagnac interferometer," *Opt. Exp.*, vol. 28, no. 5, pp. 7207–7220, Mar. 2020.
- [12] A. A. Chtcherbakov, P. L. Swart, and S. J. Spammer, "Mach-Zehnder and modified Sagnac-distributed fiber-optic impact sensor," *Appl. Opt.*, vol. 37, no. 16, pp. 3432–3437, Jun. 1998.
- [13] A. A. Chtcherbakov, P. L. Swart, S. J. Spammer, and B. M. Lacquet, "Modified Sagnac/Mach-Zehnder interferometer for distributed disturbance sensing," *Microw. Opt. Technol. Lett.*, vol. 20, no. 1, pp. 34–36, Jan. 1999.
- [14] G. Luo et al., "Distributed fiber optic perturbation locating sensor based on dual Mach-Zehnder interferometer," *Proc. SPIE*, vol. 6622, Mar. 2008, Art. no. 66220z.
- [15] D. Tu, S. Xie, Z. Jiang, and M. Zhang, "Ultra long distance distributed fiber-optic system for intrusion detection," *Proc. SPIE*, vol. 8561, Nov. 2012, Art. no. 85611W.
- [16] Q. Chen et al., "An elimination method of polarization-induced phase shift and fading in dual Mach-Zehnder interferometry disturbance sensing system," *J. Lightw. Technol.*, vol. 31, no. 19, pp. 3135–3141, Aug. 6, 2013.
- [17] C. Ma et al., "Long-range distributed fiber vibration sensor using an asymmetric dual Mach-Zehnder interferometers," *J. Lightw. Technol.*, vol. 34, no. 9, pp. 2235–2239, May 1, 2016.
- [18] J. Huang et al., "Distributed fiber-optic sensor for location based on polarization-stabilized dual-Mach-Zehnder interferometer," *Opt. Exp.*, vol. 28, no. 17, pp. 24820–24832, Aug. 2020.
- [19] H.-R. Ho, C.-Y. Hsieh, Y.-C. Hsu, and L. Wang, "Modified dual Mach-Zehnder interferometers with new locating algorithm for intrusion detection," *Opt. Exp.*, vol. 29, no. 21, pp. 34341–34359, Oct. 2021.
- [20] C. Zhang and S. Xia, "K-means clustering algorithm with improved initial center," in *Proc. 2nd Int. Workshop Knowl. Discovery Data Mining*, Jan. 2009, pp. 790–792.
- [21] M. Yedla, S. R. Pathakota, and T. M. Srinivasa, "Enhancing K-means clustering algorithm with improved initial center," *Int. J. Comput. Sci. Inf. Technol.*, vol. 1, no. 2, pp. 121–125, Jan. 2010.
- [22] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, "Efficient agglomerative hierarchical clustering," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2785–2797, Apr. 2015. [Online]. Available: <https://www.educba.com/hierarchical-clustering-analysis/>
- [23] Q. Hu, N. Ye, and M. Zhu, "Survey of cluster analysis in data mining," *Comput. Digit. Eng.*, vol. 3, no. 2, pp. 17–20, 2007.

Meng-Chen Li was born in Taiwan, in 1997. He received the B.S. degree in physics from Soochow University, New Taipei, Taiwan, in 2019, and the M.S. degree in photonics technology from National Tsing Hua University, Hsinchu, Taiwan, in 2021.

He now works with Taiwan Semiconductor Manufacturing Company, Hsinchu. His research interests include optical fiber sensors and numerical computation with computer.



Likarn Wang was born in Taiwan, in 1959. He received the B.S. and M.S. degrees in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree in electrical engineering from The Pennsylvania State University, University Park, PA, USA, in 1989.

He started working with the Department of Electrical Engineering, National Tsing Hua University, in 1990, as an Associate Professor, where he became a Full Professor in 1998. He joined the Institute of Photonics Technologies, National Tsing Hua University, in 2003. He was named the Director of the Institute in 2012 and took a three-year-and-two month term position. His research interests include optical fiber sensors, waveguide optics, and solar cells.