Optics Letters

Fiber-optic distributed seismic sensing data generator and its application for training classification nets

Lihi Shiloh,[1,*] Ariel Lellouch,[2] Raja Giryes,[1] and Avishay Eyal[1]

[1]The Electrical Engineering School at Tel-Aviv University, Tel-Aviv, Israel
[2]Department of Geophysics, Stanford University, California 94305, USA
*Corresponding author: lihishiloh@mail.tau.ac.il

Received 30 December 2019; revised 12 February 2020; accepted 18 February 2020; posted 20 February 2020 (Doc. ID 386352); published 20 March 2020

Distributed acoustic sensing (DAS) is a powerful tool thanks to its ease of use, high spatial and temporal resolution, and sensitivity. Growing demand for long-distance distributed seismic sensing (DSeiS) measurements, in conjunction with the development of efficient algorithms for data processing, has led to an increased interest in the technology from industry and academia. Machine-learning-based data processing, however, necessitates tedious *in situ* calibration experiments that require significant effort and resources. In this Letter, a geophysics-driven approach for generating synthetic DSeiS data is described, analyzed, and tested. The generated synthetic data are used to train DSeiS classification algorithms. The approach is validated by training an artificial neural-network-based classifier using synthetic data and testing its performance on experimental DSeiS records. Accuracy is greatly improved thanks to the incorporation of a geophysical model when generating training data.

https://doi.org/10.1364/OL.386352

Fiber-optic distributed acoustic sensing (DAS) is a cost-effective technology for implementing acoustic signal acquisition in large scales. In many applications, the detected signals are actually seismic waves. This version of DAS is denoted here as distributed seismic sensing (DSeiS). During recent years, DSeiS has been applied to a wide range of applications such as intrusion detection, border and perimeter defense, leak detection in pipelines, earthquake detection, transportation monitoring, vertical seismic profiling, and more [1–3]. DAS is commonly based on analysis of variations in the Rayleigh backscatter profile of a sensing fiber. Acoustic and vibration signals in the vicinity of the fiber affect the phase and amplitude of the measured profile. Extracting the amplitude variations is relatively straightforward, while obtaining reliable information from the phase measurement is more challenging. The effort is worthwhile, however, as phase signals are linearly related to the directional strain along the sensing fiber [4].

Different approaches can be used to measure backscatter profiles. The most common is the optical time domain reflectometry (OTDR). In OTDR, a short pulse is launched into the fiber, and position information is obtained from the round trip time [1]. Another approach is optical frequency domain reflectometry (OFDR). It uses chirped waveforms for interrogation, and the locations are resolved according to their beat frequencies [5]. There is no clear superior approach, and the choice between them depends on the specific application.

Regardless of the interrogation method used, all DSeiS systems require significant testing and calibration efforts following installation in a new site or seasonal/man-made changes in an existing one. One of the primary roles of the calibration phase is to collect sufficient experimental data to train machine learning detection and classification algorithms. Such efforts are typically labor-intensive and therefore expensive. Therefore, an alternative approach, utilizing the generative adversarial network (GAN) [6] was proposed. It augmented a small amount of experimental data into a sufficiently large training set [7]. However, this approach still necessitated the tedious training phase of a complex neural network architecture, the GAN, as opposed to training a common classification network.

In this study, we introduce a simple and efficient approach for generating synthetic data for the training of DSeiS detection and classification nets. Data are generated using a geophysical model with uncertainty in the subsurface parameters. The model accounts for travel time differences, amplitude decay, and directionality effects due to uniaxial strain measurement, and requires minimal computation efforts. The optical system response as a result of the seismic excitation is also applied. Without loss of generality, we assume an OFDR system. Generated data are qualitatively compared to DSeiS records acquired in field experiments.

The synthetic data is used to train an artificial neural network (ANN) classifier. The trained network is then tested on experimental recordings of footsteps, vehicle driving, and ambient noise, and achieves 92.8% accuracy (namely, 92.8% of the test set instances were correctly classified).

The method is based on a synthetic data generator which is composed of two parts: an approximated geophysical formulation of seismic wave propagation and an optical DAS interrogator, which is an OFDR in this study but can be straightforwardly adapted to an OTDR. The OFDR system was simulated similarly to the procedure described in Ref. [7]. The fiber-under-test's profile was continually generated with complex backscatter coefficients. Its profile was divided into segments of 0.5 m each, randomly drawn from a normal distribution simulating Rayleigh backscattering. Optical loss (0.2 dB/km) and noise sources, such as shot and phase noise, were added to the signal to match the experimental data characteristics. The seismic event signature, described in the following paragraph, was added as an additional phase to the fiber profile. The backscatter signal at the input of the coherent photo-receiver was obtained by convolving the fiber's impulse response with the linearly swept input waveform.

Accurately modeling seismic signatures is challenging, as they depend on source parameters and subsurface properties, which are both unknown. In addition, numerically solving the 3D seismic wave equation is computationally demanding, and building a sufficiently large training dataset would require extensive resources. Therefore, we employ a more straightforward approach. The source's signature is directly estimated from the data. Seismic sources excite mostly surface waves ($\sim 90\%$ of energy), and we thus focus on such waves in our modeling. We compute the travel times from the source location to the fiber locations using seismic surface wave velocities measured in the study area [8]. We assume a constant subsurface velocity for the propagation. Measured velocities are randomized to account for subsurface variation and errors in measurement. Signals are shifted in phase according to the computed source-receiver travel times. Due to their cylindrical propagation, surface wave amplitudes decay as one over the square root of distance in an ideal homogeneous half-space. However, to account for strong elastic scattering present in the shallow subsurface, we approximate their decay as one over the distance. In addition, we apply a constant anelastic Q-factor dissipation [9]. We only apply the frequency amplitude decay component, as formulated in the following equation:

$$H(f) = e^{-\alpha_0 f}, \qquad \alpha_0 = \frac{\pi d}{Qv}, \tag{1}$$

where d is the propagation distance, v is the surface wave velocity, and f is the frequency. Finally, fiber directivity has to be accounted for. For a surface wave recorded in a straight fiber, the strain-rate response is proportional to $vk^2 \cos^2\theta$, where k is the spatial wavenumber and θ is the angle between the propagation direction and the fiber.

The baseline source parameters were chosen according to their measured DAS seismic signatures. Footsteps were modeled as wavelets with central frequencies uniformly distributed in the range 10–20 Hz and with duration normally distributed between 10 ms and 15 ms. Synthetic vehicles' signals were created by generating white Gaussian noise and filtering it to the range 8–12 Hz, according to the power spectral density of the experimental recordings. For subsurface parameters, we used a value of $Q = 5$, which is representative of the area [10]. In order to account for lateral velocity variations, propagation to the right and left sides of the source is conducted with different velocities randomly distributed in the interval [90,166] m/s. In the experiment, the vertical distance between the fiber and the source was 10 m. For better visualization of fine features in the seismic signature, Figs. 2 and 3 show synthetic data computed for sources at a distance of 5 m from the fiber. Clearly, adaptation of the distributions of the parameters may be required for generalizing the network to different circumstances. Nonetheless, our proposed methodology provides efficient tools for such cases.

Field experiments were conducted with DSIT's Lightline OFDR interrogating system. The sensing fiber was a 700 m long fiber buried ~ 1 m underground, in an isolated field, and scanned at a repetition rate of 768 Hz. The optical data were sampled at 200 MHz, and the spatial resolution was ~ 3 m. Given the seismic velocities observed experimentally, such spatial resolution means that seismic waves with temporal frequencies > 15 Hz may be detected with some aliasing. This fact, however, is accounted for in the proposed methodology, as the same parameters of the interrogator and the ground were also used by the synthetic data generator. Field experiments included excitations, at the vicinity of the fiber, of an 80 kg walking person and a driving Renault Kangoo vehicle.

The synthetic data generator was implemented in MATLAB, and the training and testing of the NN were implemented in Python using Keras backend (TensorFlow version 1.8) with Intel core-i7 CPU, 32 GB RAM and an NVIDIA GeForce GTX 1080 Ti GPU (11 GB).

The simulation parameters of the optical interrogator were adjusted according to the OFDR system used in the field. All images presented in this Letter are waterfall representations of the seismic events. The waterfalls show measured and simulated optical differential phase, with the vertical axis as time and the horizontal axis as the distance along the sensing fiber. The differential phase waterfalls are calculated along the distance axis and filtered using FK-filters [11] for manuscript visibility only (not used for classification). Colorbars are presented in radians.

An example of the experimental seismic signatures and the corresponding synthetic generator results are presented in Fig. 1. The right column in Fig. 1, panels (b) and (d), show examples of experimental results, and the left panels (a) and (c), show synthetic generated DSeiS results. Distinctive features are observable, such as the wavefront propagation in time and distance, characteristic frequencies, amplitude decay with distance, fiber directivity, and more.

Interestingly, the synthetic generator enabled the investigation of the effects of the various subsurface parameters on the seismic signatures in the acquired DSeiS data. For this purpose,

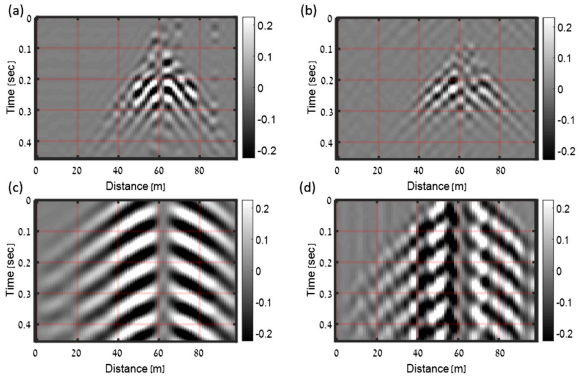

Fig. 1. Optical differential phase of seismic signatures of a footstep (first row) and a vehicle (second row). (a) and (c) are computer simulations, and (b) and (d) are experimental results.

the synthetic generator was fixed to the initial source parameters previously determined, all random noise sources were disabled, and only the subsurface parameter of interest was changed. Without loss of generality and for visibility, a footstep signature at a distance of 5 m from the fiber was used.

The effect of the subsurface velocity on the DSeiS signature is presented in Fig. 2. As can be expected, the propagating wavefront is visible, and its angle of propagation increases with the velocity. This is due to the seismic wavefronts reaching the fiber earlier. The time delay of arrival is determined by the Euclidean distance from the event source to the location along the fiber,

$$\tau = \sqrt{D^2 + x^2}/v, \tag{2}$$

where D is the vertical distance from the source to the fiber, and x is the position along the fiber (2D simulation). Notice the differences in time of arrival at a low velocity of $v = 50$ m/sec [Fig. 2(a)], and a faster one of $v = 100$ m/sec [Fig. 2(b)]. They follow the hyperbolic equation representing the Euclidean distance. Since the source location is fixed at 5 m away from the fiber, the difference in time of arrival to the closest fiber location is 46.88 msec (calculated using signal cross correlation), while the theory depicts a 50 msec difference. However, the signal's amplitude also depends on the propagation velocity. This is due to anelastic losses as well as gauge length effects. The gauge length effect can be described as an integration over different spatial locations at each time sample, equivalent to the sensor's spatial resolution. For low velocities, wavefront slopes are steep, and integration along the spatial dimension is incoherent. For high velocities, slopes are gentler, and a more coherent integration occurs. In addition, slopes also depend on location along the array—the integration is coherent near the apex of the hyperbola (zero slope). This is not the case at the tails, which tend towards a slope of x/v. This is why the hyperbolas with low velocities have frequent phase changes, whereas those with high velocities are more regular.

The effect of the Q-factor on the seismic signature is shown in Fig. 3. As the Q value decreases, the seismic signal experiences more propagation attenuation. This reinforces the suggestion that DSeiS systems deployed in soil with large Q will improve overall performances.

The influence of different event's distance from the sensing fiber is shown in Fig. 4. The event's distance decreases the signature quality (notice the different colorbar ranges) and times

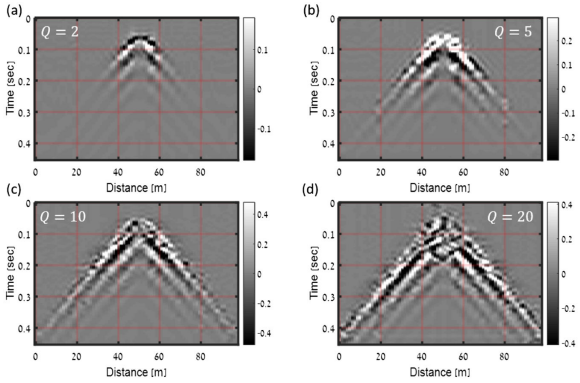

Fig. 3. Synthetic generation of a footstep. Influence of constant Q-factor dissipation on DSeiS signature in a waterfall presentation of the differential phase. (a) $Q = 2$, (b) $Q = 5$, (c) $Q = 10$, and (d) $Q = 20$.

of arrival, similar to the observations made based on Fig. 2. Another useful feature is the signal absence around the apex of the hyperbolas. As the seismic event originates at a horizontal distance from the array, the propagating wavefront reaches the closest fiber segment in a perpendicular orientation. Since particle movement of surface waves is within their plane of propagation, they do not induce strain along the direction of the fiber. As a result, there is no phase shift of the interrogating light, and there is no phase signal in the fiber segment closest to the source. As the sensitivity of the fiber is proportional to $\cos^2\theta$, the effective fiber segment that lacks sensitivity increases with source distance. This intriguing result may be used for different applications and can, for example, help locate the source relatively accurately.

To validate the similarity between synthetic and experimental data, a classification network was trained on synthetic data only and tested on experimental data. If the synthetic data include distinctive seismic features applicable to the experimental data, the network will learn them from the "free" synthetic seismic signatures and succeed in classifying the costly experimental ones.

In previous publications [7,12], an oversimplistic simulation was used to easily obtain a large training set for the classification network. This simulation did not include anelastic losses

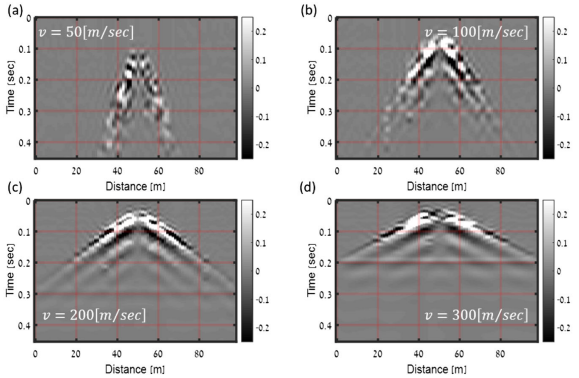

Fig. 2. Synthetic generation of a footstep. Influence of subsurface wave velocity on DSeiS signature in a waterfall presentation of the differential phase. (a) $v = 50$ m/s, (b) $v = 100$ m/s, (c) $v = 200$ m/s, and (d) $v = 300$ m/s.

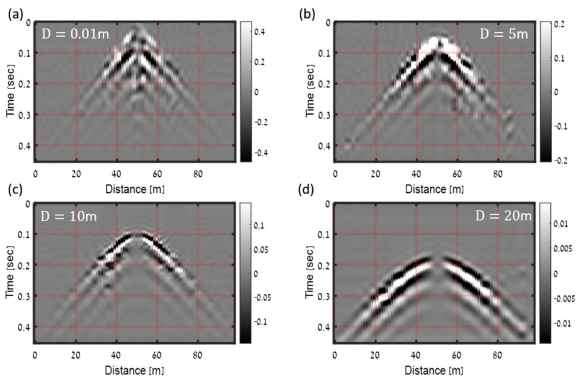

Fig. 4. Synthetic generation of a footstep. Influence of event source location from the fiber on DSeiS signature in a waterfall presentation of the differential phase. (a) $D = 0.01$ m, (b) $D = 5$ m, (c) $D = 10$ m, and (d) $D = 20$ m.

and fiber directivity, and produced random classification performance on the field experiment test data. However, it was shown that training on an NN-refined simulation data achieved good classification abilities. However, such training of a refiner network, using GAN to transform simplistic synthetic data to mimic real data, takes time and effort.

In this study, the same classification network is used but trained on the synthetically generated dataset. While it describes the seismic field more realistically, it still uses a high-frequency ray approximation of seismic wave propagation, which computes only surface Rayleigh waves, and does not account for complex subsurface structures and topography.

The network is a convolutional neural network (CNN) consisting of a modified version of Oxford's VGG16 network [13], with a two-channel input image. This input image is a concatenation of the optical power and the differential phase waterfall representation of a fiber segment (see Fig. 1). Each image is a crop of 95.5 m segment and a 0.45 s time window. This corresponds to an image of 80×350 pixels, respectively. The VGG16 convolution layers' weights are initialized by the ImageNet pretrained values, and the other layers are initialized with random weights. All layers are trained. The input images' channels were normalized, independently, to the range $[-0.5, 0.5]$. Table 1 details the full network architecture, previously referred to as FiberNet [7].

A training database was constructed of 8306 synthetic instances of each class: footsteps, vehicle driving, and ambient noise. The test set comprised 305 experimental instances from each class, manually labeled. For comparison, a similar database was constructed based on the simplistic simulation detailed in Ref. [7]. FiberNet was trained individually based on each simulation database with a learning rate of $1e-3$ to $1e-4$, data augmentation that included translation, flipping the image along the fiber-distance axis, and addition of random noise. No fine-tuning using experimental data was performed. The results are summarized in Table 2. Due to the high SNR, classification is achieved in both simulation methods. However, in the old, overly simplistic simulation, random confusion is evident between noise and vehicle. This may be attributed to the temporal continuity of the seismic signature of the vehicle combined with the lack of seismic propagation features in the simplistic simulation. Moreover, a high detection accuracy is only evident for footsteps, which indicates an overfitted state. Using the new synthetic generator, a definitive separation between all classes is achieved, and an overall increase in performance is evident for all classes with much fewer false alarms.

A detailed yet simplistic synthetic generator of seismic surface wave propagation recorded by DSeiS is described. Data generated through this scheme are representative of field DSeiS experiments, as shown by the success of a classification NN that is trained on synthetic data and tested on field data. The simulation parameters were estimated based on a small subset of the experimental data. Classification between ambient noise, footsteps, and driving vehicle was achieved with a 92.8% accuracy. This result proves that the presented seismic model can be used for a first-order analysis of DSeiS systems and training classification networks with minimal expenses of effort and money.

Table 1. FiberNet Architecture

Layer	Size In	Size Out
Conv2D	$350 \times 80 \times 2$	$350 \times 80 \times 3$
VGG16 (13 conv. layers with max pooling)	$350 \times 80 \times 3$	$20 \times 1 \times 512$
FC1 (ReLU + BN)	$1 \times 10{,}240$	1×4096
FC2 (ReLU + BN)	1×4096	1×4096
FC3 (ReLU + BN)	1×4096	1×128
Logits (softmax)	1×128	1×3

Table 2. Classification Accuracy for a 5 km Fiber and Three Classes over One Training Session (Noise, Footsteps, and Vehicles)

Training Setup	Accuracy	Confusion Matrix		
Old simulation dataset	68.6%	**58.3%**	3%	38.7%
		0%	**94.7%**	5.3%
		46.6%	3.6%	**52.8%**
New synthetic generator dataset	**92.8%**	**93.4%**	0%	6.6%
		4.63%	**91.1%**	4.6%
		5.9%	0.3%	**93.8%**

Funding. European Research Council [757497 (SPADE)]; Israel Innovation Authority; Shlomo Shmeltzer Institute for Smart Transportation at Tel-Aviv University.

Acknowledgment. The authors would like to thank NVIDIA for the GPU grant and Yossi Sharon from DSIT Solutions Ltd. for the experimental data and helpful discussions. The first author is partially supported by a scholarship from the Shlomo Shmeltzer Institute for Smart Transportation at Tel-Aviv University.

Disclosures. AV: DSIT Solutions Ltd (C).

REFERENCES

1. J. Tejedor, J. Macias-Guarasa, H. F. Martins, D. Piote, J. Pastor-Graells, S. Martin-Lopez, P. Corredera, and M. Gonzalez-Herraez, Sensors **17**, 355 (2017).
2. A. Lellouch, S. Yuan, Z. Spica, B. Biondi, and W. L. Ellsworth, JGR Solid Earth **124**, 6931 (2019).
3. A. Masoudi, M. Belal, and T. P. Newson, Meas. Sci. Technol. **24**, 085204 (2013).
4. H. Gabai and A. Eyal, Opt. Lett. **41**, 5648 (2016).
5. D. Arbel and A. Eyal, Opt. Express **22**, 8823 (2014).
6. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds. (Curran Associates, 2014), Vol. **27**, pp. 2672–2680.
7. L. Shiloh, A. Eyal, and R. Giryes, J. Lightwave Technol. **37**, 4755 (2019).
8. A. Lellouch and M. Reshef, Geophysics **82**, S9 (2017).
9. E. Kjartansson, J. Geophys. Res. **84**, 4737 (1979).
10. M. Shustak and A. Lellouch, Geophysics **83**, R699 (2018).
11. O. Yilmaz, *Seismic Data Analysis*, 2nd ed. (Society of Exploration Geophysicists, 2012).
12. L. Shiloh, A. Eyal, and R. Giryes, in *26th International Conference on Optical Fiber Sensors*, OSA Technical Digest (Optical Society of America, 2018) paper ThE22.
13. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, *ICLR* (2015), pp. 1–14.