Sound Event Classification meets Data Assimilation with Distributed Fiber-Optic Sensing

Noriyuki Tonami[1], *Yoshiyuki Yajima*[1], *Wataru Kohno*[2], *Sakiko Mishima*[1], *Reishi Kondo*[1], *Tomoyuki Hino*[1]

[1]NEC Corporation [2]NEC Laboratories America, Inc.

***Abstract*—Distributed Fiber-Optic Sensing (DFOS) is a promising technique for large-scale acoustic monitoring. However, its wide variation in installation environments and sensor characteristics causes spatial heterogeneity. This heterogeneity makes it difficult to collect representative training data. It also degrades the generalization ability of learning-based models, such as fine-tuning methods, under a limited amount of training data. To address this, we formulate Sound Event Classification (SEC) as "data assimilation" in an embedding space. Instead of training models, we infer sound event classes by combining pretrained audio embeddings with simulated DFOS signals. Simulated DFOS signals are generated by applying various frequency responses and noise patterns to microphone data, which allows for diverse prior modeling of DFOS conditions. Our method achieves out-of-domain (OOD) robust classification without requiring model training. The proposed method achieved accuracy improvements of 6.42, 14.11, and 3.47 percentage points compared with conventional zero-shot and two types of fine-tune methods, respectively. By employing the simulator in the framework of data assimilation, the proposed method also enables precise estimation of physical parameters from observed DFOS signals.**

***Index Terms*—sound event classification, distributed fiber-optic sensing, data assimilation**

1. INTRODUCTION

Distributed fiber-optic sensing (DFOS), also known as coherent optical time-domain reflectometry (C-OTDR) or ϕ-OTDR [1], [2], is an emerging sensing technique that enables the detection of sound and vibration signals using standard optical fibers. A key advantage of DFOS is its ability to monitor large-scale environments with fine spatial resolution. This results in densely sampled multichannel data over long distances. Such large-scale and high-resolution sensing capabilities have led to a wide range of applications, including whale call detection in the ocean [3], traffic monitoring [4], seismic activity observation [5], sound event recognition [6], long-range monitoring over distances exceeding 100 km [7], [8], and utility pole localization in urban areas [9]. These examples demonstrate the versatility of DFOS as a distributed acoustic sensing platform and its potential for diverse sound analysis tasks.

Sound Event Classification (SEC) [10], [11] is a task of recognizing sound event classes, e.g., "dog," "car," and "people speaking" from audio. In sound event classification, most studies have explored deep-neural-network (DNN)-based classification with microphone [10], [12], [13]. With large-scale monitoring enabled by DFOS, applications for sound event classification are expected to become even more active.

The domain gap between microphone and DFOS signals is a major challenge in implementing DFOS-driven SEC [6]. While transfer learning or fine-tuning microphone-pretrained models on DFOS data can improve performance, these methods are inherently limited by the large diversity within the DFOS domain itself. DFOS systems vary widely in both their installation environments, such as underwater, underground, and overhead locations, and in their sensor characteristics, such as frequency response and optical noise sensitivity. The combination of environmental diversity and variability in sensor properties creates a large number of possible sensing conditions, many of which are not covered during training. As a result, even when a model is adapted to a specific DFOS setup, it may not generalize to other scenarios. Since any learning-based method optimizes its parameters based on the training distribution, it tends to perform poorly when applied to different inference conditions. This makes it especially difficult to ensure robust generalization in practice, particularly when collecting labeled DFOS data for every condition is unrealistic.

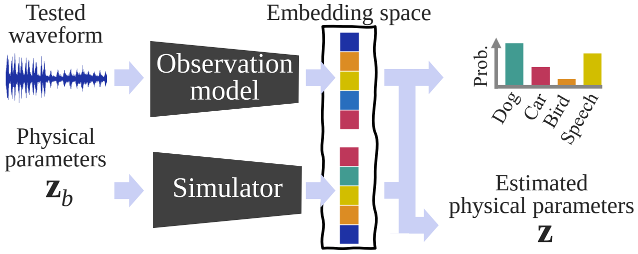

Fig. 1: Concept of proposed SEC with data assimilation

To address the limitations of training-based adaptation under domain mismatch, we reformulate SEC as "data assimilation" in an embedding space, as shown in figure 1. Instead of updating model parameters, our method estimates the most plausible class by combining pretrained audio representations with simulated or observed DFOS signals. We simulate DFOS responses across diverse scenarios to account for spatial heterogeneity and map them to the same embedding space using a pretrained audio encoder. At inference time, we identify the optimal class by considering posteriors of observed and simulated DFOS signals in the shared embedding space, without retraining or fine-tuning. This approach enables domain-robust classification, even under severe data scarcity and heterogeneous sensing conditions.

2. PRELIMINARIES

2.1. Principle of DFOS

In DFOS, the phase of the backscattered light generated at every location along the fiber cable is observed:

$$\Delta\mathbf{\Phi} = [\Delta\Phi_1, \ldots, \Delta\Phi_c, \ldots, \Delta\Phi_C] \in \mathbb{R}^{C\times T} \quad (1)$$

where $\Delta\Phi_c = [\Delta\phi_{c,1}, \ldots, \Delta\phi_{c,t}, \ldots, \Delta\phi_{c,T}]^\top$ indicates a waveform signal at location c. C and T denote the number of the observed locations, i.e., channels, along the fiber and the total duration. Given a local acoustic pressure $\epsilon_{x,t}$ at position x along the fiber cable and time t, $\Delta\phi_{c,t}$ is approximated as follows:

$$\Delta\phi_{c,t} \propto \underbrace{\int_{-L/2}^{L/2} \epsilon_{x,t} dx}_{\text{low-pass filter}} \quad (2)$$

$L \in \mathbb{R}^+$ is known as gauge length, which is a hyperparameter of DFOS systems. The point is that $\Delta\phi_{c,t}$ is proportional to the summation of the local acoustic pressure at time t and within position c to $c + L$. In other words, Eq. 2 is regarded as a directional moving average filter [14] which is one of the low-pass filters.

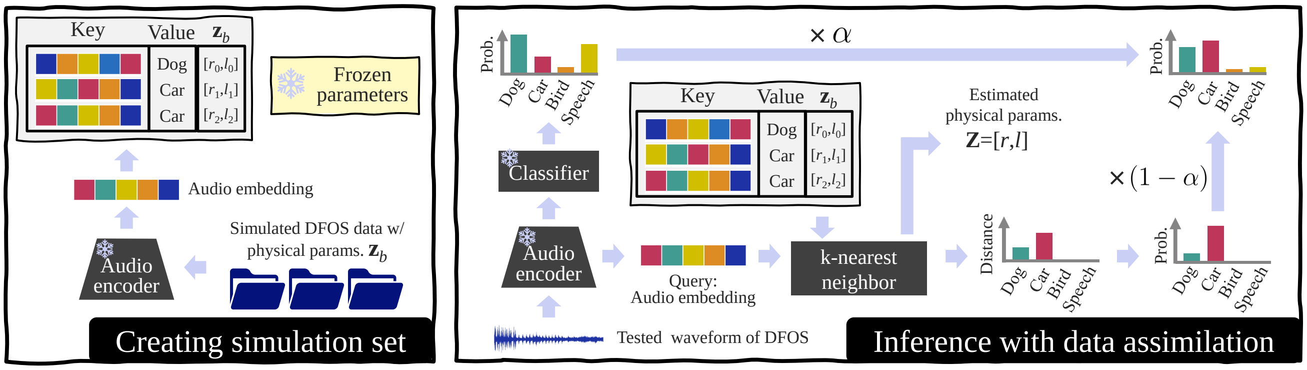

Fig. 2: Overview of proposed SEC with data assimilation

Given a frequency response and noise, the DFOS waveform $\Delta\Phi_c$ at channel c is also expressed as:

$$\Delta\widetilde{\Phi}_c = \Delta\Phi_c * \boldsymbol{\mathcal{H}}_c + \mathbf{n}_c, \tag{3}$$

where $*$ denotes convolution, and $\boldsymbol{\mathcal{H}}_c$ and $\mathbf{n}_c$ represent the frequency response and additive noise at channel c, respectively. These channel-dependent characteristics arise from the effect of Eq. 2 and diverse installation environments such as seabeds, underground, and aerial deployments. Moreover, $\mathbf{n}_c$ has an effect on optical noise, which follows a Gaussian distribution. Due to the large number of channels C, there exist many combinations of $\boldsymbol{\mathcal{H}}_c$ and $\mathbf{n}_c$. It is impractical to collect all these variations for statistical modeling. This spatial heterogeneity leads to a significant domain gap between training and testing. As a result, model adaptation, such as finetuning, becomes difficult, particularly when the amount of training data is limited.

2.2. Data assimilation

Data assimilation is a technique for improving predictions, including physical parameters, by combining simulation with new observations. It has been applied in fields such as weather forecasting [15], [16] and traffic monitoring [4], where simulation-based predictions are adjusted using observations. By leveraging simulation-based priors, data assimilation offers robustness even under out-of-domain conditions, where real observations deviate from training or modeling assumptions. Data assimilation is especially useful when the available data are limited, noisy, or differ from the conditions assumed in the model.

In data assimilation, the errors of simulations and observations are minimized to obtain estimations from a simulator and its physical parameters. Let $\mathbf{z}_b \in \mathbb{R}^n$ be the background (prior) state estimate, and $\mathbf{z}_o \in \mathbb{R}^m$ be the observation vector. The observation operator $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}^m$ maps the state space to the observation space. Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ be the background error covariance matrix and $\mathbf{R} \in \mathbb{R}^{m\times m}$ be the observation error covariance matrix. The widely used 3D-Var cost function [15], [17], [18], which is the core expression of data assimilation, is defined as:

$$J(\mathbf{z}) = \underbrace{\frac{1}{2}\left(\mathbf{z}-\mathbf{z}_b\right)^T \mathbf{B}^{-1}\left(\mathbf{z}-\mathbf{z}_b\right)}_{\text{simulation error}} + \underbrace{\frac{1}{2}\left(\mathcal{H}(\mathbf{z})-\mathbf{z}_o\right)^T \mathbf{R}^{-1}\left(\mathcal{H}(\mathbf{z})-\mathbf{z}_o\right)}_{\text{observation error}} . \tag{4}$$

The optimal analysis state $\mathbf{z}$, which is the physics parameters, is obtained by minimizing the cost function:

$$\mathbf{z} \leftarrow \arg\min_{\mathbf{z}} J(\mathbf{z}) . \tag{5}$$

3. PROPOSED METHOD: DATA ASSIMILATION-BASED INFERENCE FOR DFOS

We treat SEC as a data assimilation problem to address the spatial heterogeneity in the DFOS-based SEC with a limited amount of real DFOS training data. Instead of training a model using DFOS data, we estimate the sound event class by comparing pretrained audio embeddings with observed or simulated DFOS signals. This approach allows out-of-domain (OOD) robust classification without retraining and estimates physical parameters using a simulator as a white-box approach.

3.1. Formulation of data Assimilation in embedding space

We formulate data assimilation in an embedding space where both simulation and observed signals are mapped through a pretrained model f. By sharing embedding space for both the simulation and observed signals, we can assume that the observation operator $\mathcal{H}$ is the identity function. Furthermore, assuming decorrelation of the embedding vectors, we approximate the error covariances $\mathbf{B}$ and $\mathbf{R}$ with identity matrices. Under these assumptions, the cost function Eq. 4 of the data assimilation simplifies to:

$$J(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}-\mathbf{z}_b\|_2^2 + \frac{1}{2}\|\mathbf{z}-\mathbf{z}_o\|_2^2 . \tag{6}$$

In our method, we employ k nearest neighbor (kNN) model in an embedding space to obtain the optimal state of $\mathbf{z}$.

3.2. Simulated DFOS in embedding space

To implement data assimilation of SEC under the concept of Eq. 6, we generate a simulation dataset $\mathcal{D}_{\text{simDFOS}}$, which is a set of key-value pairs with physics parameters $\mathbf{z}$, by applying various DFOS transfer functions, i.e., various low-pass filters [14] and noise, to microphone-domain audio x_i.

$$(\mathcal{K}, \mathcal{V}, \mathcal{Z}) = \{(f_{\text{mic}}(x_i^{(\text{sim})}), y_i, \mathbf{z}_b^{(i)}) \mid (x_i^{(\text{sim})}, y_i, \mathbf{z}_b^{(i)}) \in \mathcal{D}_{\text{simDFOS}}\}, \tag{7}$$

where $x_i^{(\text{sim})}$ denotes the simulated DFOS signal of class y_i. $f_{\text{mic}} : \mathbb{R}^T \to \mathbb{R}^D$ denote a pretrained audio encoder that maps a T-dimensional waveform to a D-dimensional feature vector. For the simulation of DFOS signals, we consider the cutoff frequency r of low-pass filters and signal-to-noise ratio (SNR) l of Gaussian noise as in Eqs. 2 and 3: $\mathbf{z}_b^{(i)} = [r_i, l_i]$.

3.3. Posterior estimation utilizing data assimilation

In inference stages, given a query, i.e., an audio embedding $f_{\text{mic}}(x^{(\text{test})})$ of a tested DFOS waveform $x^{(\text{test})}$, the kNN-based model retrieves k nearest neighbor key-value pairs of the simulation

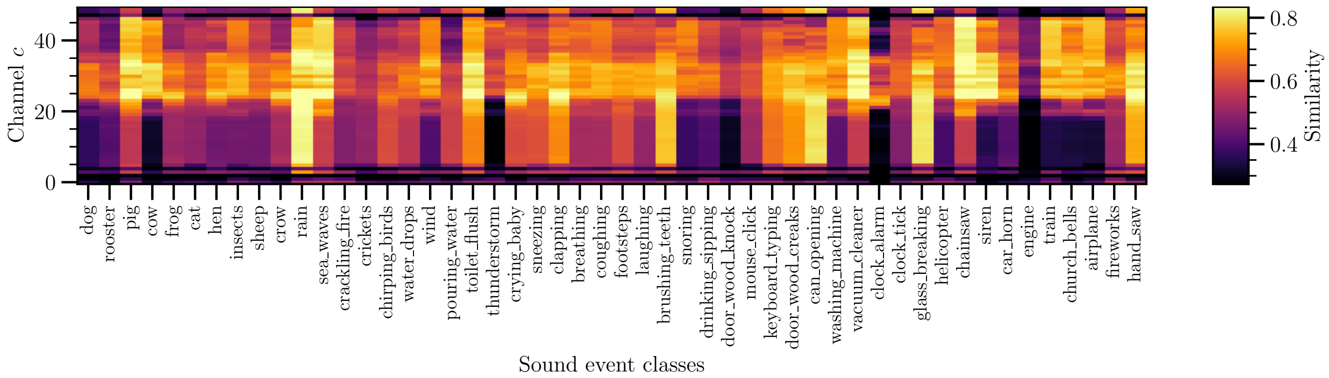

Fig. 4: Similarity in frequency response between DFOS & microphone

set $\mathcal{D}_{\mathrm{simDFOS}}$. Given the set of k nearest neighbor key-value pairs $\mathcal{N}$, the simulation probability is

$$p_{\mathrm{sim}}(y|x) = \frac{1}{|\mathcal{N}|} \sum_{(k_i, v_i) \in \mathcal{N}} \mathbb{1}_{y=v_i} \exp\Big(-d\big(k_i, f_{\mathrm{mic}}(x^{(\mathrm{test})})\big)\Big), \tag{8}$$

where $\mathbb{1}_{y=v_i}$ denotes an indicator function that takes one when i-th value v_i is equal to an estimated event label y; otherwise zero. $d(\cdot,\cdot)$ indicates a distance function, which is squared L^2 distance in our work.

In parallel, we obtain the classifier-based prediction as observation probability using only microphone-domain models:

$$p_{\mathrm{obs}}(y|x) = \mathsf{softmax}\Big(f_{\mathrm{clf}}\big(f_{\mathrm{mic}}(x^{(\mathrm{test})})\big)\Big)_y, \tag{9}$$

where $f_{\mathrm{clf}} : \mathbb{R}^D \to \mathbb{R}^E$ is a pretrained classifier that maps the embedding to E-dimensional logits over event classes. $\mathsf{softmax}(\cdot)$ indicates the softmax function over sound event classes, and then the subscript y selects the probability for class y.

Finally, as a solution of Eq. 6, we interpolate the probability $p_{\mathrm{sim}}(y|x)$ with the output of the observation probability $p_{\mathrm{obs}}(y|x)$ to obtain the final prediction:

$$p(y|x) = \alpha p_{\mathrm{obs}}(y|x) + (1-\alpha) p_{\mathrm{sim}}(y|x), \tag{10}$$

where $\alpha \in [0, 1]$ balances the obserbation- and simulation-based priors. This formulation enables OOD robust inference under spatially diverse DFOS conditions without requiring model adaptation or retraining.

3.4. Estimation of physical parameters via data assimilation

The physical parameters $\mathbf{z} = [r, l]$, which are the lowcut frequency r of the low-pass filters and SNR l of gaussian noise, are also estimated based on data assimilation using k nearest neighbor set $\mathcal{N}$ of Eq. 8:

$$r \leftarrow \frac{1}{|\mathcal{N}|} \sum_{\mathbf{z}_b^{(i)} \in \mathcal{N}} \mathrm{proj}_r(\mathbf{z}_b^{(i)}), \quad l \leftarrow \frac{1}{|\mathcal{N}|} \sum_{\mathbf{z}_b^{(i)} \in \mathcal{N}} \mathrm{proj}_l(\mathbf{z}_b^{(i)}), \tag{11}$$

where $\mathrm{proj}_r(\cdot)$ and $\mathrm{proj}_l(\cdot)$ represent an operator of extracting an element r and l of vector, respectively.

4. EXPERIMENTS

4.1. Experimental conditions

Recording: To validate the proposed method, we first conduct a recording of DFOS data. As shown in Fig. 3, the optical fiber is embedded into a mat. The figure's left and right sides depict the real picture and its illustration, respectively. The size of the mat is 1.2 m × 1.2 m. Sound source signals were played using a speaker and then omnidirectionally propagated to the optical fiber of the mat. The straight-line distance between the edge of the mat and the speaker is about 2.8 m. We then obtained 50 channels from the DFOS data of the mat. For the sound sources, we used the ESC-50 dataset [10], which is comprised of 2,000 audio clips. The sampling frequency of DFOS is 20 kHz. Gauge length L is set to 4 m. Figure 4 shows the similarity in the frequency domain between original dry sources (ESC-50) and recorded DFOS data. Cosine similarity was used for the similarity.

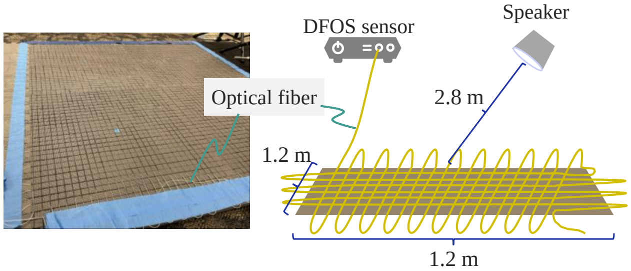

Fig. 3: Installation of optical fiber

Simulation procedure of DFOS: DFOS exhibits lower sensitivity at higher frequencies [14], [19] and considerable variability across different sensing locations. Based on those works, we simulated DFOS data from the ESC-50 dataset [10] with low-pass filters and Gaussian noise to reproduce Eq. 3. For the low-pass filters, we used 4th-order Butterworth low-pass filters with cutoff frequencies r {1, 2, 3, 4, 5, 6, 7} kHz. Gaussian noise is then added to the low-pass waveform of ESC-50 with SNR l {-5, -10, -15} dB. We finally generated 42,000 simulated DFOS waveforms from the ESC-50 dataset. The generated data was used for $\mathcal{D}_{\mathrm{simDFOS}}$ and in a manner of cross-validation.

Evaluation setting: We conducted a five-fold cross-validation following [10] with the real DFOS dataset. As shown in figure 4, DFOS data has the spatial heterogeneity. As $f_{\mathrm{mic}}(x)$, we used hierarchical token semantic audio Transformer (HTS-AT) [20] from Contrastive language-audio pretraining [21], referred to as "MS-CLAP." For MS-CLAP, the sound events of ESC-50 are classified by prompts "*this is the sound of [class label]*," in accordance with [22] using generative pretrained Transformer 2 (GPT2) [23] as $f_{\mathrm{cls}}(x)$. For kNN, k was set to 50 tuned using the folds of the training. For comparison, we used the following methods.

- Zero-shot: zero-shot classification based on MS-CLAP was used.
- Fine-tune: linear probe [24] was used, referred to as "fine-tune." In the linear probe, HTS-AT with fully connected layers was trained using Adam optimizer [25].

4.2. Experimental results

4.2.1. Spatial heterogeneity: We explore the classification performance under the spatial heterogeneity in DFOS. For clarity, we set

Table 1: Accuracy [%] of channels $c = 25 - 50$ for condition of spatial heterogeneity

Method	Types of data for training		α	DFOS channels of observed data used for finetuning pretrained model											
				Out-of-domain (0–20)					In-domain (20–50)						
	Observed	Simulated		0–5	5–10	10–15	15–20	Avr.	20–25	25–30	30–35	35–40	40–45	45–50	Avr.
Zero-shot	No training		1.0	18.33	18.33	18.33	18.33	18.33	18.33	18.33	18.33	18.33	18.33	18.33	18.33
Fine-tune	✓		–	3.11	9.71	4.85	**24.90**	10.64	58.30	49.94	63.18	64.10	59.80	56.59	58.65
		✓	–	21.28	21.28	21.28	21.28	21.28	21.28	21.28	21.28	21.28	21.28	21.28	21.28
Proposed	No training		0.0	**24.75**	**24.75**	**24.75**	24.75	**24.75**	24.75	24.75	24.75	24.75	24.75	24.75	24.75

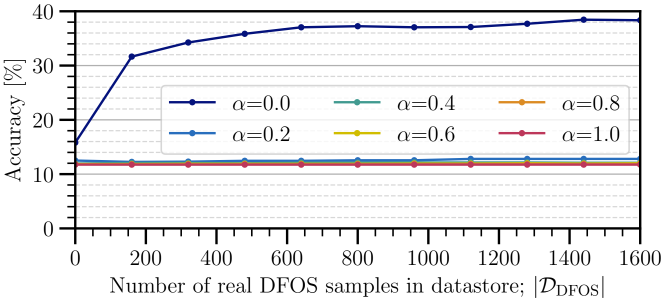

Fig. 5: Impact of real DFOS samples $\mathcal{D}_{\mathrm{DFOS}}$ and α

$\alpha = 0.0$ in Eq. 10 of the proposed method to verify the effectiveness of the simulation set. As shown in figure 4, DFOS data has a large diversity in the frequency response of DFOS compared with that of the microphone. This spatial heterogeneity is caused by a directional filter and optical noise of Eqs. 2 and 3. As can be seen in the figure, the 20th channel and after have high similarities between DFOS and the microphone; otherwise, there are low similarities. In this experiment, we divide the DFOS channels into two sections, before and after the 20th, to verify their performance under the condition of spatial heterogeneity.

Table 1 shows the accuracy of classifying events where channels $c = 25 - 50$ were used for the inference. For the fine-tuning method, we used two types of data: observed and simulated data. The observed data means the real DFOS data captured by the fiber mat in Fig. 3. The result indicates that the proposed method outperforms the conventional methods in terms of average accuracy under $c = 0 - 20$; OOD section. The proposed method improved the classification performance by 6.42, 14.11, and 3.47 percentage points in terms of average accuracy compared with the zero-shot and fine-tune methods with the observed or simulated DFOS data.

The proposed method achieves significantly better performance than the conventional methods for each channel of $c =$ 0–15. In particular, the proposed method with $\mathcal{D}_{\mathrm{simDFOS}}$ improved the accuracy by 21.64 percentage points compared with the fine-tuning methods. On the other hand, the performance of the conventional fine-tuning method gets drastically worse. This is because the lower-SNR DFOS data cause the fine-tuning method to catastrophically forget the acoustic features of the pretrained model.

For $c =$ 20–50, the conventional fine-tuning method with the observed DFOS data outperformed the other methods, including the proposed methods. Since there is a small gap between the tested and trained channels compared with those of $c =$ 0–20, the fine-tuning method with the observed DFOS data works well.

4.2.2. Impact of number of real samples and α in data assimilation: In this experiment, we relaxed the constraint by using actual DFOS data for the proposed data assimilation SEC. Specifically, we add a real DFOS dataset $\mathcal{D}_{\mathrm{DFOS}}$, which is used for the fine-tune methods, into the simulated DFOS dataset $\mathcal{D}_{\mathrm{simDFOS}}$. The real DFOS dataset $\mathcal{D}_{\mathrm{DFOS}}$ was constructed using the same procedure as $\mathcal{D}_{\mathrm{simDFOS}}$.

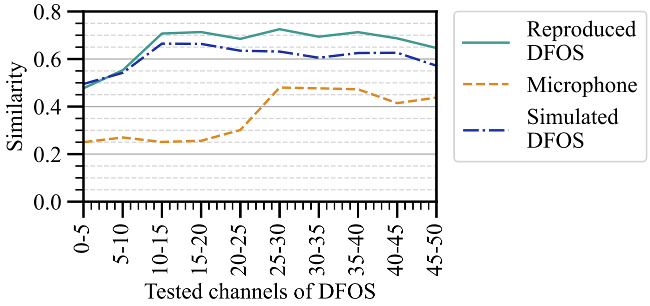

Fig. 6: Similarities with tested real DFOS data

Figure 5 shows the all-channel-averaged accuracy with changing $\mathcal{D}_{\mathrm{DFOS}}$ and α. The result shows that the accuracy of the proposed method with $\alpha =$ 0.2–1.0 is significantly lower than that of $\alpha = 0.0$. This indicates that the large-scale models trained by the microphone are helpful for encoding audio not observed by the microphone, even though there is a large gap between DFOS and the microphone. When $\alpha = 0.0$, we found that increasing the size of $\mathcal{D}_{\mathrm{DFOS}}$ leads to improved performance, achieving high accuracy even with a small number of samples.

4.2.3. Estimation of physical parameters: Figure 6 shows the performance of estimating the physical parameters $\mathbf{z}$ with Eq. 11. In the figure, the green line depicts the average of the cosine similarities in the embedding space of MS-CLAP between tested DFOS signals and reproduced DFOS signals. The reproduced DFOS signals were generated from the original microphone signal of the ESC-50 dataset with the estimated physical parameters $\mathbf{z} = [r, l]$. The orange dashed line represents the average cosine similarity in the MS-CLAP embedding space between the tested DFOS signals and the original microphone signals from the ESC-50 dataset. The blue dotted line shows the average similarity between the tested DFOS signals and the simulated DFOS signals. The result shows that the similarities between the tested DFOS signals and the reproduced DFOS signals are higher than those of the other methods. In other words, the proposed data-assimilation-based SEC precisely estimates the physical parameters $\mathbf{z} = [r, l]$. The result indicates that the proposed method estimates the optimal physical parameters by interpolating the prior physical parameters $\mathbf{z}_b$ of the simulation.

5. CONCLUSION

We addressed the challenge of the spatial heterogeneity in the DFOS-based SEC under a limited amount of real DFOS training data. Our method formulates inference as data assimilation in the embedding space, combining pretrained audio features with simulated DFOS signals, and avoids model retraining. The proposed method achieved accuracy improvements of 6.42, 14.11, and 3.47 percentage points compared with zero-shot and two fine-tune methods. The proposed data-assimilation-based SEC also precisely estimates the physical parameters of a tested DFOS signal compared to the simulated DFOS signals.

REFERENCES

[1] E. Ip, Y. Huang, M. Huang, M. Salemi, Y. Li, T. Wang, Y. Aono, G. Wellbrock, and T. Xia, "Distributed fiber sensor network using telecom cables as sensing media: Applications," *Proc. Optical Fiber Communications Conference and Exhibition* (*OFC*), pp. 1–3, 2021.

[2] E. Ip, J. Fang, Y. Li, Q. Wang, M. Huang, M. Salemi, and Y. Huang, "Distributed fiber sensor network using telecom cables as sensing media: technology advancements and applications," *Journal of Optical Communications and Networking* (*JOCN*), vol. 14, no. 1, pp. 61–68, 2022.

[3] L. Bouffaut, K. Taweesintananon, H. Kriesell, R. Rorstadbotnen, J. Potter, M. Landro, S. Johansen, J. Brenne, A. Haukanes, O. Schjelderup, and F. Storvik, "Eavesdropping at the speed of light: Distributed acoustic sensing of baleen whales in the arctic," *Frontiers in Marine Science*, vol. 9, pp. 1–13, 2022.

[4] Y. Yajima, H. Prasad, D. Ikefuji, T. Suzuki, S. Tominaga, H. Sakurai, and M. Otani, "A novel approach to real-time short-term traffic prediction based on distributed fiber-optic sensing and data assimilation with a stochastic cell-automata model," *Proc. the Transportation Research Board Annual Meeting* (*TRB*), pp. 1–22, 2025.

[5] T. Parker, S. Shatalin, and M. Farhadiroushan, "Distributed acoustic sensing a new tool for seismic applications," *First Break*, vol. 32, no. 2, pp. 61–69, 2014.

[6] W. Kohno, N. Tonami, J. Fang, S. Han, J. Sun, and T. Wang, "Text-guided device-realistic sound generation for fiber-based sound event classification," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 1–5, 2025.

[7] O. Waagaard, E. Rønnekleiv, A. Haukanes, F. Stabo-Eeg, D. Thingbø, S. Forbord, S. Aasen, and J. Brenne, "Real-time low noise distributed acoustic sensing in 171km low loss fiber," *OSA Continuum*, vol. 4, no. 2, pp. 688–701, 2021.

[8] E. Rønnekleiv, T. Sørgård, D. Klimentov, N. Tolstik, O. Waagaard, J. Jacobsen, F. Stabo-Eeg, O. Sab, A. Calsat, P. Plantady, and J. Brenne, "Range-scalable distributed acoustic sensing with edfa repeaters demonstrated over 2227 km," *Optics Letters*, , no. 1, pp. 25–28, 2025.

[9] Y. Lu, Y. Tian, S. Han, E. Cosatto, S. Ozharar, and Y. Ding, "Automatic fine-grained localization of utility pole landmarks on distributed acoustic sensing traces based on bilinear resnets," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4675–4679, 2021.

[10] J. Piczak, "ESC: Dataset for environmental sound classification," *Proc. the 23rd Annual ACM Conference on Multimedia* (*ACMM*), pp. 1015–1018, 2015.

[11] H. Sailor, D. Agrawal, and H. Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification," *Proc. INTERSPEECH*, pp. 3107–3111, 2017.

[12] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with CNN-Transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (*TASLP*), vol. 28, pp. 2450–2460, 2020.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. International Conference on Learning Representations* (*ICLR*), pp. 1–21, 2021.

[14] T. Dean, T. Cuny, and A. Hartog, "The effect of gauge length on axially incident p-waves measured using fibre optic distributed vibration sensing," *Geophysical Prospecting*, vol. 65, pp. 184–193, 2016.

[15] D. M. Barker, W. Huang, Y-R. Guo, A. J. Bourgeois, and Q. N. Xiao, "A three-dimensional variational data assimilation system for mm5: Implementation and initial results," *Monthly Weather Review*, vol. 132, no. 4, pp. 897–914, 2004.

[16] M. Leutbecher and T.N. Palmer, "Ensemble forecasting," *Journal of Computational Physics*, vol. 227, no. 7, pp. 3515–3539, 2008.

[17] S. Dobricic and N. Pinardi, "An oceanographic three-dimensional variational data assimilation scheme," *Ocean Modelling*, vol. 22, no. 3, pp. 89–105, 2008.

[18] B. Melinc and Z. Zaplotnik, "3D-Var data assimilation using a variational autoencoder," *Quarterly Journal of the Royal Meteorological Society*, vol. 150, no. 761, pp. 2273—2295, 2024.

[19] N. Tonami, W. Kohno, S. Mishima, Y. Arai, R. Kondo, and T. Hino, "Low-rank constrained multichannel signal denoising considering channel-dependent sensitivity inspired by self-supervised learning for optical fiber sensing," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 8511–8515, 2024.

[20] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 646–650, 2022.

[21] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 336–340, 2024.

[22] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: learning audio concepts from natural language supervision," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 1–5, 2023.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *Tech. Rep., OpenAI*, 2019.

[24] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *Proc. International conference on machine learning* (*ICML*), pp. 8748–8763, 2021.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. International Conference on Learning Representations* (*ICLR*), 2015.