LEARNABLE WAVELET PACKET TRANSFORM FOR DATA-ADAPTED SPECTROGRAMS

Frusque Gaëtan and Fink Olga

ETH Zurich, Swiss Federal Institute of Technology, Zurich, Switzerland

ABSTRACT

Capturing high-frequency data concerning the condition of complex systems, e.g. by acoustic monitoring, has become increasingly prevalent. Such high-frequency signals typically contain time dependencies ranging over different time scales and different types of cyclic behaviors. Processing such signals requires careful feature engineering, particularly the extraction of meaningful time-frequency features. This can be time-consuming and the performance is often dependent on the choice of parameters. To address these limitations, we propose a deep learning framework for learnable wavelet packet transforms, enabling to learn features automatically from data and optimise them with respect to the defined objective function. The learned features can be represented as a spectrogram, containing the important time-frequency information of the dataset. We evaluate the properties and performance of the proposed approach by evaluating its improved spectral leakage and by applying it to an anomaly detection task for acoustic monitoring.

Index Terms— Wavelet Packet Transform, Sparse Decomposition, High Frequency Signals, Anomaly Detection, Harmonic Analysis.

1 Introduction

Capturing high-frequency data concerning the condition of complex systems, e.g. by acoustic monitoring, has become increasingly prevalent. As a first step, learning tasks based on high-frequency signals typically require an extraction of relevant time-frequency features before they are used as inputs in machine learning (ML) algorithms. ML algorithms are generally designed for inputs of a fixed size. Fourier and wavelet transforms provide very good feature extraction capabilities and are able to reduce the dimensionality of the data. However, their application requires time-consuming engineering of the features that is dependent on both the dataset and the problem.

Learning the features automatically from raw data can help to overcome the challenges of manual feature engineering. Deep learning (DL) methods like Convolutional Neural Networks (CNN) are able to automatically learn representative features from raw signals [1], [2]. However, their performance on time-series data is highly dependent on the hyperparameters and the architecture. Both of these are typically specific to the dataset. Furthermore, the resulting algorithms are not easily interpreted by the domain experts. Recently, several architectures that combine the advantages of signal processing and the learning capabilities of neural networks have shown promising results. Some examples include successful applications on speaker recognition with kernels that are constrained to correspond to band-pass filters only [3], sound classification with a learned filter-bank [4], molecular partners prediction, and utilisation of a learned wavelet transform in the field of cosmology [5]. These models turned out to be more interpretable or to contain fewer parameters than a standard CNN.

In this work, we propose a deep learning architecture, based on wavelet packet transform (WPT), called L-WPT. Our approach provides an automatically learnt sparse time-frequency representation of a univariate signal. While the previously proposed learnable wavelet transform approaches [6], [7], [8], [5] focus on learning the best filter to use throughout the entire architecture, in this research, we seek to generalise the filter learning and add learned biases for automatic denoising. While a similar approach was recently proposed in [9] in the context of discrete wavelet transforms (DWT), the advantage of the WPT structure over DWT is that it decomposes the signal in spectral bands of similar width and with a similar time resolution, enabling a better spectral resolution on the high frequencies. Consequently, L-WPT can be seen as an ideal tool to learn data-adapted spectrograms, concurrently reducing the number of parameters in the neural network significantly. We demonstrate that the proposed approach is capable of mitigating the spectral leakage problem that is inherent to WPT. We further show the applicability of the proposed architecture for detecting anomalies in acoustic monitoring dataset.

This study was supported by the Swiss Innovation Agency (lnnosuisse) under grant number: 47231.1 IP-ENG.

2 Background

2.1 Wavelet Packet Transform (WPT)

The discrete WPT, introduced in [10], projects the signal on uniform frequency bands of desired size. The WPT has a multi-level structure and can be considered a multiresolution analysis since the output of the current level is recursively used as input to the next level. The basic bloc of a WPT applied to an input signal $\mathbf{y}$ is:

$$(\mathbf{y}^{lp})_{(n)} = (\mathbf{h}^{lp} * \mathbf{y})_{(2n)}, \quad (\mathbf{y}^{hp})_{(n)} = (\mathbf{h}^{hp} * \mathbf{y})_{(2n)}. \quad (1)$$

where $*$ is the convolution operation and $\mathbf{y}^{lp}$ ($\mathbf{y}^{hp}$) corresponds to the low- (high-) pass filtered input data with a cut-off frequency of $\pi/2$. This transformation doubles the frequency resolution (the frequency content of each wavelet coefficient spans half the input data frequency) to the detriment of a halved time resolution ($\mathbf{y}^{lp}$ and $\mathbf{y}^{hp}$ each contains half the number of samples in $\mathbf{y}$). Recursively, we use $\mathbf{y}^{lp}$ and $\mathbf{y}^{hp}$ as input of the same bloc eq. (1) to generate frequency bands of desired width. This process gives as many outputs (also called nodes) as 2^L, where L is the number of recursions (called layers).

2.2 Perfect reconstruction

It is possible to recover the input signal $\mathbf{y}$ from $\mathbf{y}^{lp}$ and $\mathbf{y}^{hp}$ via inverse WPT (iWPT). The following constraints on the filters are sufficient to ensure a perfect reconstruction [11]; we separate them into relational constraints between filters and constraint on the filter's coefficients.

- **Architectural properties**: $\mathbf{h}^{hp}$ has to be the Alternating Flip (AF) of $\mathbf{h}^{lp}$. It causes both filters to have mirrored magnitude responses around $\pi/2$. The AF of a filter $\mathbf{h}$ is denoted as flip[$\mathbf{h}$] and corresponds, for a filter of size $K+1$ with K odd and $k \in \{0, ..., K\}$, to:

$$\text{flip}[\mathbf{h}]_{(k)} = (-1)^k (\mathbf{h})_{(K-k)}. \quad (2)$$

Also, the iWPT filters have to respect the Anti-Cancellation Conditions (ACC) to remove aliasing terms due to the use of non-ideal filters. This can be done by considering the delayed paraconjugate of $\mathbf{h}^{hp}$ and $\mathbf{h}^{lp}$. Applied to a real filter $\mathbf{h}$, this operation results in $\bar{\mathbf{h}}$ with:

$$(\bar{\mathbf{h}})_{(k)} = (\mathbf{h})_{(K-k)}. \quad (3)$$

Notice that by respecting both AF and ACC properties, we only have to design one filter $\mathbf{h}^{lp}$.

- **Kernel property**: to ensure the conservation of the frequency content, $\mathbf{h}^{hp}$ and $\mathbf{h}^{lp}$ have to be power complementary (the power sum of both filters has to be one for all frequencies). Considering $\mathbf{h}^{hp}$ is the AF of $\mathbf{h}^{lp}$, then, these filters are conjugate mirrors [12], which implies that $\mathbf{h}^{lp}$ coefficients have to be normalised and double-shift orthogonal (for non-null shifts).

2.3 Denoising with the WPT

Signal denoising is one of the major applications of wavelet analysis [13]. It has been shown that, concerning regular and structured signals, a wavelet transform will lead to a sparse decomposition [12]. We can then assume that the noise will correspond to wavelet coefficients of small amplitudes. To eliminate these low-amplitude wavelet coefficients, we consider the sharp, hard-thresholding activation function proposed in [14] and denoted as HT[x]:

$$\text{HT}[x; \gamma] = x(\sigma_{(-10(x+\gamma))} + \sigma_{(10(x-\gamma))}), \quad (4)$$

with $\sigma_{(x)} = 1/(1+e^{-x})$ the sigmoid function and γ the bias acting as thresholds on both sides of the origin.

3 Method: Learned-WPT

The proposed method is a learnable deep architecture based on WPT and referred to as Learned-WPT (L-WPT). This section details the proposed methodology.

3.1 Architecture of L-WPT

L-WPT resembles the WPT structure. However, the filter elements for each bloc and each layer are learned independently. The activation function eq. (4) is applied to the output of each filter. We impose the filters to respect all the architectural properties presented in section 2.2. This has the advantage of reducing the number of filters that need to be learnt by a factor of four, to keep the mirrored frequency interpretation and the ACC property. However, the kernel property can't be preserved, which will prevent the perfect reconstruction of the input signal.

By applying L-WPT, the wavelet coefficients at layer l, node i (denoted $\mathbf{y}_l^i$) are convoluted with the filter θ_l^i to produce the wavelet coefficients of the upper layer:

$$(\mathbf{y}_{l+1}^{2i})_{(n)} = \text{HT}[(\theta_l^i * \mathbf{y}_l^i)_{(2n)}; \gamma_{l+1}^{2i}], \quad (5)$$

$$(\mathbf{y}_{l+1}^{2i+1})_{(n)} = \text{HT}[(\text{flip}[\theta_l^i] * \mathbf{y}_l^i)_{(2n)}; \gamma_{l+1}^{2i+1}], \quad (6)$$

It is simply a rewriting of eq. (1), considering a different filter for each bloc, the AF eq. (2) and the learned activation function eq. (4). Similar to the iWPT, it is possible to estimate the wavelet coefficients $\mathbf{y}_l^i$ from the upper layer via inverse L-WPT. The computation of the estimated reconstruction of $\mathbf{y}_l^i$, considering eq. (3) to respect the ACC property, is denoted $\hat{\mathbf{y}}_l^i$ and is written:

$$\hat{\mathbf{y}}_l^i = \overline{\theta_{\mathrm{l}}^{\mathbf{i}}} * \text{up}[\hat{\mathbf{y}}_{l+1}^{2i}] + \overline{\text{flip}[\theta_{\mathrm{l}}^{\mathbf{i}}]} * \text{up}[\hat{\mathbf{y}}_{l+1}^{2i+1}] \quad (7)$$

The upsampling operation up[$\bullet$] is necessary to counteract the effect of the downsampling in eq. (5); it can be defined as $\text{up}[\mathbf{x}]_{(2n)} = x_{(n)}$ and $\text{up}[\mathbf{x}]_{(2n+1)} = 0$.

Fig. 1 illustrates the decomposition and reconstruction via L-WPT showing the decomposition tree for the case of $L = 2$ layers.

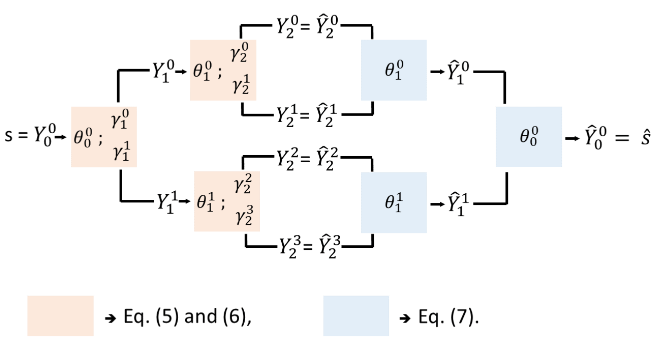

$s = Y_0^0 \rightarrow \theta_0^0$; γ_1^0 γ_1^1
$Y_1^0 \rightarrow \theta_1^0$; γ_2^0 γ_2^1
$Y_1^1 \rightarrow \theta_1^1$; γ_2^2 γ_2^3
$Y_2^0 = \hat{Y}_2^0$
$Y_2^1 = \hat{Y}_2^1$
$Y_2^2 = \hat{Y}_2^2$
$Y_2^3 = \hat{Y}_2^3$
θ_1^0
θ_1^1
$\rightarrow \hat{Y}_1^0$
$\rightarrow \hat{Y}_1^1$
θ_0^0
$\rightarrow \hat{Y}_0^0 = \hat{s}$
→ Eq. (5) and (6),
→ Eq. (7).

Fig. 1: Decomposition tree of L-WPT with $L = 2$

3.2 Loss for data-adapted spectrogram

We refer to $\mathbf{Y}_L = [\mathbf{y}_L^0, \mathbf{y}_L^1, ..., \mathbf{y}_L^{2^L-1}]$ as the set of all node outputs for the last layer L. The matrix $|\mathbf{Y}_L|$ can be considered a learned spectrogram representing the temporal evolution of 2^L frequency bands. We propose to impose sparsity on the obtained spectrogram. We aim to learn the best representation of the signal with the filters $\boldsymbol{\Theta} = [\theta_0^0, \theta_1^0, \theta_1^1, ..., \theta_{L-1}^{2^{L-1}-1}]$ as high-amplitude wavelet coefficients that are sparse. The activation function eq. (4) with learned bias $\boldsymbol{\Gamma} = [\gamma_1^0, \gamma_1^1, ..., \gamma_L^{2^L-1}]$ will suppress the low wavelet coefficient associated with noise and induce sparsity in $|\mathbf{Y}_L|$. Inspired by [14], the proposed loss function to be minimised is defined as:

$$\underset{\boldsymbol{\Theta},\boldsymbol{\Gamma}}{\operatorname{argmin}} \quad || \, \mathbf{s} - \hat{\mathbf{s}} \, ||_1 + \alpha \, || \, \mathbf{Y}_L \, ||_1, \tag{8}$$

where $\mathbf{s} = \mathbf{y}_0^0$ is the input data and $\hat{\mathbf{s}} = \hat{\mathbf{y}}_0^0$ is the reconstructed signal. $|| \bullet ||_1$ is the ℓ_1-norm function and corresponds to a surrogate of the non-convex ℓ_0-norm to induce sparsity. α is the trade-off parameter between the sparsity of the spectrogram and the quality of the reconstruction. The ℓ_1-norm is used for both the residual and the regularisation to ease the selection of the trade-off parameter. Indeed, by choosing $\alpha = 1$, we give the same importance to both parts of the eq. (8) since $\mathbf{Y}_L$ and $\mathbf{s}$ contain the same number of elements.

3.3 Learning strategy

The proposed framework can be implemented as a deep auto-encoding architecture. The encoding blocs correspond to two CNN layers (one per filter) with a shared kernel, stride 2 for the downsampling and followed by the activation functions eq. (4). The decoding blocs contain the corresponding transposed CNN layers with upsampling. Considering L layers and $K+1$ coefficients per filter, the number of trainable parameters is $\sum_{l=1}^{L} 2^{(L-l)}K$ for the filters and $\sum_{l=1}^{L} 2^L$ for the bias.

4 Evaluation case studies

For both applications, we use WPT and L-WPT with $L = 5$ layers, which involves an embedding $\mathbf{Y}_L$ of 32 nodes (now denoted $\mathbf{Y}$ since we no longer consider the previous layers). We initialise the filters of L-WPT using the Daubechies wavelets with 8 coefficients (denoted db4), and the biases are initialised to 0.3. The L-WPT is trained using the Adam optimiser with a learning rate of 0.001.

4.1 Spectral leakage analysis

Using non-ideal filters in WPT generates a spectral leakage and makes the discrimination of harmonic content imperfect. To evaluate the performance of our framework, we propose, based on [15], to measure the level of distortion produced by L-WPT. For this experiment, we generate a 10-seconds Frequency-Swept Cosine (FSC), with a sampling frequency of $f_s = 2^{13}$, which sweeps from null to the maximum frequency $f_s/2$ such as:

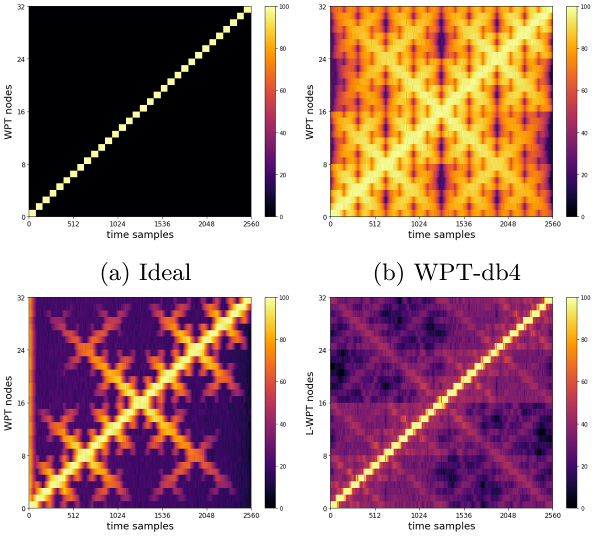
$$w_\beta(n) = \cos(\pi 0.05 n^2) + \beta \mathcal{N}_{(0,1)}, \tag{9}$$

where $n \in \{1, ..., 10f_s\}$, $\mathcal{N}_{(0,1)}$ is a realisation of a normal distribution and β is the noise level. The ideal WPT (based on ideal filters) of the pure FSC ($\mathbf{w}_{\beta=0}$) consists of each node successively activating each of the 80 samples according to the sweep level (see fig. 2(a)). To assess the quality of the transform, we propose computing the Root Sum Square (RSS) of the difference between the estimated representation and the ideal one. Both representations are first normalised to have the same l2-norm.

To learn the coefficients and biases of L-WPT, we generate 400,000 noisy cosinus signals with a random frequency. Each can be written as $s(n) = \cos(2\pi f n) + \mathcal{N}_{(0,1)}$, where f is a realisation of a uniform distribution taking values from 0 to $f_s/2$.

The time-frequency representation of the signal $\mathbf{w}$ using L-WPT is compared to the results of WPT with different wavelet families or filters. We consider the standard Daubechies wavelet with 8 and 46 coefficients (WPT-db4 and WPT-db23). We also consider the best filters for measuring the fluctuating harmonics using WPT, which are, according to [15]: Butterworth with 12 coefficients (WPT-butt) and Elliptic with 8 coefficients (WPT-ellip). In contrast to the Daubechies wavelet, these are Infinite Impulse Response (IIR) filters, which should improve the frequency selectivity. For the three last filters, we have selected the number of coefficients (in the range from 2 to 76) which best minimise the RSS score.

Figs. 2 (b), (c), and (d) show the spectral leakage for WPT-db4, WPT-ellip, and L-WPT in decibels (DB). We can observe the consequent spectral leakage introduced by the commonly used WPT-db4. The spectral discrimination is improved with WPT-butt. However, it seems to be impacted by border effects and delays. Finally, the spectrogram obtained with L-WPT is close to the ideal one. Fig. 3 shows the mean RSS score over 100 realisations of $\mathbf{w}_\beta$ for $\beta \in \{0, 0.1, 0.2, ...2\}$. For all noise levels, L-WPT outperforms all other methods. For pure FSC ($\beta = 0$), the RSS score for the L-WPT is 0.49 and 0.53 for the WPT-butt (second-best method). Our method is robust to noise. The biggest gap is obtained at $\beta = 0.6$, where the RSS score for L-WPT is 0.54 and 0.85 for WPT-butt.



(a) Ideal

(b) WPT-db4

(c) WPT-butt

(d) L-WPT

Fig. 2: Output of the pure FSC $\mathbf{w}_{\beta=0}$ through WPT with different filters (a), (b), (c), and our L-WPT (d). The colormap is based on the % of the maximum value.

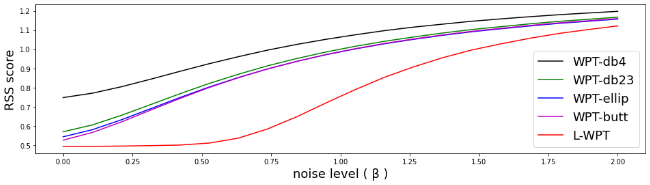


Fig. 3: RSS score for WPT with different filters and L-WPT in function of the noise level β

4.2 Anomaly detection with L-WPT

We evaluate the performance of our model on an anomaly detection task with the Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection (MIMII). It consists of audio recordings of four types of industrial machines, *i.e.*, valves, pumps, fans, and slide rails. The training and validation dataset is set as in [16]. The training dataset contains sound recordings from healthy machines only. We aim to detect anomalies without any knowledge of the fault characteristic.

Machine	Fan	Pump	Slider	Valve	All
WPT	91.3	85.6	97.8	97.0	92.9
DesPawn	92.4	82.2	92.5	94.0	90.3
WPT-CNN	91.8	86.8	98.5	**97.3**	93.6
L-WPT	**94.9**	**87.7**	**99.5**	97.0	**94.8**

Table 1: mean AUC for the different machine types

After learning a data-adapted spectrogram representation of the healthy sounds via L-WTP, a 1-class Extreme Learning Machine (ELM) with 100 neurons [14] is applied for fault discrimination with the input as the mean and maximum of the residual and of each frequency band (following [9]).

We compare the embedding learned by the L-WPT with a) the standard WPT-db4, b) the learned DWT with $L = 17$ layers (called DesPawn) [14], and c) the architecture of the WPT-CNN from [6] (contrary to the L-WPT, it does not have an activation function and the same learned filter is used throughout the entire structure). For each data-driven method, the results may vary with the number of epochs and the hyperparameter α. We fix $\alpha = 1$ for the L-WPT and DesPawn and provide the best area under the curve (AUC) over 500 epochs resulting from the classification of the 1-class ELM.

Table 1 shows the AUC scores for each machine and each method. DesPawn is outperformed by the WPT methods (except for the fan case). This demonstrates the importance of having a better spectral resolution at high frequencies for this application. The L-WPT gives the best results overall, outperforming the WPT by 1.9% and the learned WPT-CNN by 1.2%. It outperforms other approaches on each machine except for the valve, where the WPT-CNN yields a slightly better result. Here, the level of sparsity α seems to be too high.

5 Conclusion

We propose a new deep architecture that integrates the learning abilities of deep learning with the advantages of signal processing approaches. The proposed L-WPT is able to learn a data-adapted spectrogram of the signal. It achieves superior performance in discerning pure harmonic content and outperforms other similar methods on an anomaly detection task for acoustic monitoring.

There are several ways to extend the proposed framework, such as integrating additional constraints into the objective function, analysing the interpretability of the proposed framework, and extending it to multivariate signals.

6 References

[1] Wei Zhang, Gaoliang Peng, Chuanhao Li, Yuanhang Chen, and Zhujun Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, pp. 425, 2017.

[2] Jonathan J Huang and Juan Jose Alvarado Leanos, "Aclnet: efficient end-to-end audio classification cnn," *arXiv preprint arXiv:1811.06669*, 2018.

[3] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[4] Hardik B Sailor, Dharmesh M Agrawal, and Hemant A Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification.," in *Interspeech*, 2017, vol. 8, p. 9.

[5] Wooseok Ha, Chandan Singh, Francois Lanusse, Eli Song, Song Dang, Kangmin He, Srigokul Upadhyayula, and Bin Yu, "Adaptive wavelet distillation from neural networks through interpretations," *arXiv preprint arXiv:2107.09145*, 2021.

[6] Shoucong Xiong, Hongdi Zhou, Shuai He, Leilei Zhang, Qi Xia, Jianping Xuan, and Tielin Shi, "A novel end-to-end fault diagnosis approach for rolling bearings by integrating wavelet packet transform into convolutional neural network structures," *Sensors*, vol. 20, no. 17, pp. 4965, 2020.

[7] Daniel Recoskie and Richard Mann, "Learning sparse wavelet representations," *arXiv preprint arXiv:1802.02961*, 2018.

[8] Dhruv Jawali, Abhishek Kumar, and Chandra Sekhar Seelamantula, "A learning approach for wavelet design," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5018–5022.

[9] Gabriel Michau and Olga Fink, "Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series," *arXiv preprint arXiv:2105.00899*, 2021.

[10] Ronald R Coifman, Yves Meyer, and Victor Wickerhauser, "Wavelet analysis and signal processing," in *In Wavelets and their applications*. Citeseer, 1992.

[11] Gilbert Strang and Truong Nguyen, *Wavelets and filter banks*, SIAM, 1996.

[12] Stéphane Mallat, *A wavelet tour of signal processing*, Elsevier, 1999.

[13] Fábio M Bayer, Alice J Kozakevicius, and Renato J Cintra, "An iterative wavelet threshold for signal denoising," *Signal Processing*, vol. 162, pp. 10–20, 2019.

[14] Gabriel Michau, Yang Hu, Thomas Palmé, and Olga Fink, "Feature learning for fault detection in high-dimensional condition monitoring signals," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 234, no. 1, pp. 104–115, 2020.

[15] Jorge Bruna and Julio J Melero, "Selection of the most suitable decomposition filter for the measurement of fluctuating harmonics," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 11, pp. 2587–2594, 2016.

[16] Harsh Purohit, Ryo Tanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi, "Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, p. 209.