WaveNet based architectures for denoising periodic discontinuous signals and application to friction signals

Jules Rio, Fabien Momey, Christophe Ducottet, Olivier Alata
Laboratoire Hubert Curien, CNRS UMR 5516, Université Jean Monnet, IOGS
Université de Lyon
42023 Saint-Étienne, France

***Abstract*—In this paper, we introduce a deep learning model based on Wavenet to denoise periodic signals containing some strong discontinuities, where the dataset used for training contains only synthetic data. We introduce a new cost function using a total variation term. The synthetic data which contain strong discontinuities, are generated as the sum of a sine wave, a square signal and a white gaussian noise. This simple model is very time-efficient to compute and allows us to perform data generation for each training of the architecture instead of physically storing the dataset. We specifically apply this model to real friction signals obtained through a rotating tribological system. We also compared our method with an improved TV denoising algorithm.**

***Index Terms*—Denoising, Wavenet, Total variation, Synthetic data, Friction signals**

I. Introduction

The denoising task, which consists of extracting a clean signal $\mathbf{x}$ from a mixture $\mathbf{y} = \mathbf{x} + \mathbf{n}$ with $\mathbf{n}$ a noise, has been widely explored in the last decades. It is often a prerequisite to other tasks such as fault diagnosis [1], [2] or automatic speech recognition [3].

In this paper, we investigate the case of periodic signals with strong discontinuities (an example is given in Fig. 1(*Left*)). This kind of signals can be encountered in a wide range of applications such as audio-signals, electrocardiograms and rotating machinery. We particularly deal with friction signals obtained through a rotating tribological system. Methods based on empirical mode decomposition have also been proposed with application to friction signals [4].

As conventional methods [5] usually do not work well with discontinuities, total variation (TV) based methods have received a lot of attention [6]–[8]. However, these methods tend to create a solution defined as a staircase function. [9] proposed to couple TV-denoising with wavelets in order to overcome this constraint.

For audio signals, conventional techniques [5] have been widely replaced by deep learning techniques, such as convolutional networks [10]–[12], generative adversarial networks [13]–[15] or LSTM [3], thanks to their ability to introduce both regularity constraints and data based priors. Audio source separation, another task close to audio enhancement, has also been widely studied with deep learning techniques [16]–[18]. These methods are generally supervised or semi-supervised, and therefore require a large amount of clean data, which is not always available, especially for friction data.

To overcome the lack of training data, a possibility is to use synthetic data. The use of synthetic data is not new and was already proposed by [19], [20] for image applications or by [21], [22] for video applications. However, the creation of their data rely on a database of existing objects. Unlike them, we propose here to create the training data on the fly, allowing us to reduce the physical storage (our training set would require around 3GB). Additionally, we propose to use an architecture closely related to the "WaveNet for speech denoising" introduced by [12]. However, it contains much less parameters and can therefore be stored more easily, and can process much longer signals.

In this paper, we introduce a denoising algorithm based on a deep learning model trained with specific synthetic data. Our contribution is to propose (i) a WaveNet based deep learning model adapted to periodic signals having discontinuities, (ii) a training process relying on specific synthetic data and (iii) a regularization loss to improve generalization.

In Section II, we recall the main aspects of [12]. In Section III, we present the specificities of our work. Evaluations and applications are shown in Section IV.

II. Wavenet for speech denoising

The wavenet for speech denoising [12] is based on [23], which performed audio generation in an end-to-end manner, avoiding to discard information contained in the phase. Both architectures are a succession of stacks of residual layers returning a residual connection and a skip connection. All skip connections are summed and processed by two convolutional layers, both preceded by a *ReLU* (*Rectified Linear Unit*). Residual layers are gated units where the contribution of a dilated convolution (with kernel size of 3×1), with a tanh activation, is controlled by a sigmoid, and 1×1 convolutions define skip and residual connections. Each stack is made of L residual layers with dilation factors $1, 2, \ldots, 2^{L-1}$, which

This work was supported by the IMOTEP project within the 'Investissements d'Avenir' program operated by the French Environment and Energy Management Agency (ADEME)

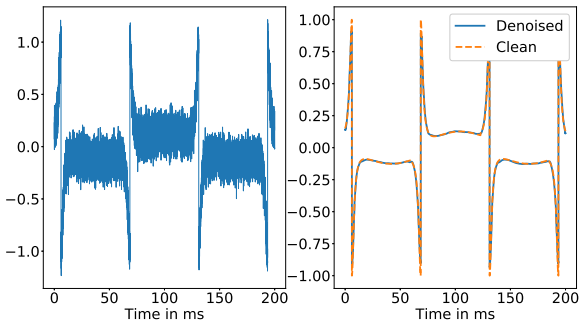


Fig. 1: Signal generated with a friction model
Left: Noisy - *Right*: Denoised and ground truth

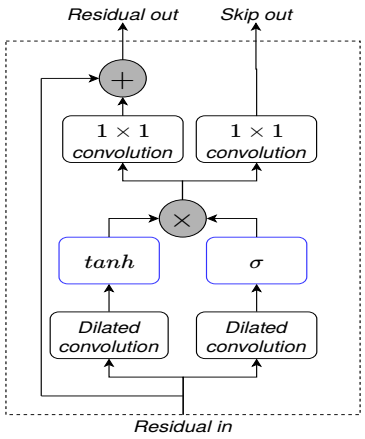


Fig. 2: Details of a residual layer

efficiently increases the receptive field. There are S successive stacks, allowing a deeper architecture without increasing the receptive field as much as with a single stack of $S \times L$ layers. 1×1 convolutions are used at the beginning and end of the network to get the adequate amount of channels.

As [12] is used for denoising, it also has some specificities: all convolutions are non-causal and time continuity is ensured by using 3×1 kernels in the two convolutions performed after summing skip connections, rather than using an autoregressive process as in [23] (which is non-parallelizable and therefore time-consuming).

Regarding details of the architecture, [12] used $L = 10$ and $S = 3$, with 128 channels in the residual layers. The two convolutions after summing skip connections have 2048 and 256 filters respectively. This configuration will be referred as "Original" in the next sections.

III. Method : A WaveNet for discontinuous periodic signals

A. Leveraging of the WaveNet for denoising

In this paper, we adapt the architecture presented in section II for our application. The different configuration are shown in Table I.

TABLE I: Reduced architectures details

	Light-Full	**Medium**	**Small**
Residual layers per stack	10	8	6
Stacks	3	2	2
Residual layers : channels	64	64	64
Channels in the last layers	256,64	256,32	256,32
Parameters (Original : 6.3M)	1.1M	600K	460K

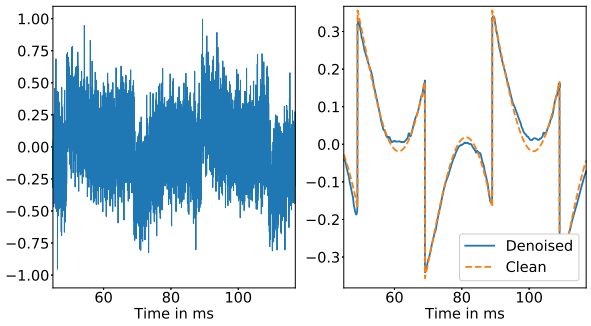


Fig. 3: Signal generated with the synthetic model
Left: Noisy - *Right*: Denoised and ground truth

Although the depth of the network and its receptive field are decreased (except for the "Light-Full" architecture), this shortened version of the architecture allows the processing of much longer signals thanks to the fully convolutionnal structure and the fewer memory used to store the model. The smaller number of parameters also requires less memory to be stored and reduces the risks of overfitting.

B. Loss

For training the network, mean absolute error (MAE, the $L1$ loss divided by the amount of samples, which was preferred to $L2$ based on [12] and better generalization in our first experiments) was combined with a total variation term whose purpose is to promote solutions having slow varying parts.

$$\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T}|\hat{\mathbf{x}}_t - \mathbf{x}_t| + \frac{\gamma}{T}\sum_{t=2}^{T}|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t-1}| \quad (1)$$

where $\mathbf{x}$ is the clean signal, $\hat{\mathbf{x}}$ is the retrieved signal, T is the amount of samples in the signal and γ is a weight to balance the two terms.

A similar regularization was proposed in [24] and [25] for image processing.

C. Training

As previously introduced, our WaveNet-based denoiser is dedicated to smooth periodic signals presenting outlying or periodic discontinuities. Hence, training signals are generated as a composition of a square wave (discontinuous component) and a sine wave. For each generated signal, the first component is randomly dephased with the second. This results in signals $\mathbf{y} = \{\mathbf{y}_t\}_{t\in\mathbb{Z}}$ where:

$$\mathbf{y}_t = \lambda c(2\pi f t + \phi_0) + (1-\lambda)\sin(2\pi f t + \phi_1) + \mathbf{n}_t \quad (2)$$
$$= \mathbf{x}_t + \mathbf{n}_t \quad (3)$$

where c is the square wave, $(\phi_0, \phi_1) \in [0, 2\pi] \times [0, 2\pi]$, f is a frequency in the set $\{0.25, 0.5, 1, 2, 4, 8, 16, 24, 32\}$Hz, shared by the two components, λ is a random weight in $[0, 1]$ and $\mathbf{n}$ is a white gaussian noise with signal-to-noise ratio (SNR) in the set $\{-10, -5, 0, 5, 10, 20\}$dB where:

$$SNR = 10 \log_{10} \left(\frac{\|\mathbf{x}\|_2^2}{\|\mathbf{n}\|_2^2} \right) \quad (4)$$

As a weighted sum of a sine wave and a square wave, the ground truth lays in range [-1,1]. An example of these signals is shown on Figure 3 (*Left*). To avoid overfitting discontinuities, 50% of training data were the sum of a constant and a white noise. Using noise-only data was shown efficient by [12]. We use a larger amount of these data because we find it useful for better generalization in our case.

The obtained dataset contained 10000 signals, each containing 20000 samples. Additionally, 3000 signals of 20000 samples were created with the same method (without noise-only data) to perform validation.

IV. Experiments

A. Experimental setup

We trained the model on 30 epochs with an initial learning rate of 0.001, divided by 10 after 20 epochs. The model was trained using Adam optimizer and batches of 8 signals. The best model was obtained based on validation score.

B. Comparison with the full wavenet

Evaluation was performed on two synthetic sets:

Synthetic set Signals in this set are generated with the same synthetic model as the one used for training (without noise-only data) with frequencies in $\{10, 20, 25, 35, 40\}$Hz and SNR randomly picked in $\{-15, -7, -3, 3\}$dB.

Friction set Signals in this set are friction signals based on the following model defined in [26] :

$$\mu(V) = \left(\mu_c + (\mu_s - \mu_c)\, e^{-\left(\frac{|V|}{v_s}\right)} \right) \times \frac{V}{|V|} + k_s V \quad (5)$$

where μ_c, μ_s, i and k_s are constants and V is a periodic speed. Noise is added to $\mu(V)$ with the same SNR as in training set.

For these two datasets, 3000 signals of 20000 samples were generated with a sample rate of 10^5Hz.

As the value of the source of interest in the training signals is always in $[-1, 1]$, all evaluation signals were divided by their maximum absolute values. While it would be possible to perform this normalization with ground truth values, values considered were those of the noisy signal in order to check the ability of the model to generalize to real contexts where ground truth is unknown. The signals are brought back to their initial range after denoising.

An example of denoising result obtained for the synthetic set can be seen on Figure 3 and an example of the friction model set can be seen on Figure 1 (using medium architecture trained with $\gamma = 10$). These two examples show that the trained model managed to efficiently fit the data.

TABLE II: Results on synthetic set (SNR)

	Original	**Light-Full**	**Medium**	**Small**
$\gamma = 0$	17.35	15.70	**18.02**	16.35
$\gamma = 1$	16.68	15.48	17.93	16.45
$\gamma = 10$	16.22	14.46	17.29	15.89
$\gamma = 100$	-0.03	12.46	15.87	15.03

TABLE III: Results with friction model (SNR)

	Original	**Light-Full**	**Medium**	**Small**
$\gamma = 0$	25.98	23.34	25.17	22.91
$\gamma = 1$	25.17	24.06	**25.38**	22.98
$\gamma = 10$	23.44	23.25	24.88	22.41
$\gamma = 100$	6.26	18.49	22.16	20.68

To have a fair idea of the performances of our architecture, we compared it with the full version as described in Section II, trained on 15 epochs with an initial learning rate of 0.001 divided by 10 after 10 epochs (longer training did not give significant improvements), with a training set defined the same way as the one used for our model. The training signals had 8000 samples instead of 20000 due to memory constraints. To get the same amount of samples, 25000 were used. For the evaluation, the same sets of 3000 signals as for our models were used. To be able to process the full 20000 samples, each signal is cut in three segments and the full prediction is obtained by concatenating the three predictions.

We evaluated the results with SNR. The comparisons can be seen in Table II for the synthetic set and in Table III for the friction set (refer to Table I for details on architectures, "Original" stands for the configuration of [12]). Comparing the three architectures, we can remark that the "Medium" one performs better or almost equally as the "Original" one. Also note that the "Light full" and the "Small" ones are just a bit lower. This confirms that reducing the architecture is not damaging. Concerning the influence of the TV regularization, the value $\gamma = 100$ gives poorer results due the degradation of peaks in the signal. We also notice that the best results are obtained without regularization in the synthetic set and with a slight regularization in the friction set. This tends to motivate the use the TV loss for better generalization but it has to be confirmed with additional experiments (see Section IV-D).

C. Comparison with an improved TV denoising algorithm

Reference [9] also proposed a method for denoising discontinuous signals, where a total variation is combined with the use of wavelets. Denoising is performed with the "Watv_software"[1] they provide (with the same parameters for the algorithm as proposed in their paper) on five of the 1s signals they use (*Piece-Regular* (**PR**), *Piece-Polynomial* (**PP**), *Ramp* (**R**), *HeaviSine* (**HS**) and Blocks (**B**)) with 1024 samples. While these signals are not periodic, they still contain strong discontinuities, which is the main aspect of our work.

To perform prediction with our model, which requires signals in range $[-1, 1]$, signals were centered then divided by their maximum absolute value and the inverse process was

[1] http://eeweb.poly.edu/iselesni/software/index.html

TABLE IV: Results (in SNR) with the method of [9]

Original SNR	PR	PP	R	HS	B
-15dB	0.97	0.85	1.38	1.45	0.80
-5dB	8.94	8.84	10.64	11.24	8.48
5dB	15.89	16.40	20.38	20.13	15.66
15dB	25.10	26.42	30.23	28.48	28.88

TABLE V: Results (in SNR) with our method, improvements over [9] are reported in bold

Original SNR	PR	PP	R	HS	B
-15dB	**2.04**	**2.35**	**4.6**	**6.26**	**3.42**
-5dB	6.86	8.25	**14.24**	**12**	7.01
5dB	13.48	15.1	**25.43**	**20.55**	14.04
15dB	15.3	19.54	**34.04**	**29.33**	16.54

applied after denoising. A white noise was added to reach a given SNR. As our previous model was not made to fit a sample rate of 1024Hz, we retrained our architectures with a training set of signals of 4096 samples with this sample rate (and where the frequencies of the synthetic signals were the original ones that were lower than 4Hz). We used $\gamma = 1$, which generally gave the best results for this application. In the proposed signals, a large number of discontinuities may appear within a few samples. Our models are not made to deal with a lot of discontinuities in the receptive field. Therefore, only the results of the "Small" architecture, which gave the best results, are shown.

We made four of these experiments, with four different original SNR : -15dB, -5dB, 5dB, 15dB. The denoising results are evaluated with SNR of the denoised signal.

Tables IV and V) show the results of respectively the method of [9] and our model on the signals (all values are obtained as the average of 100 realisations of noise). Our model seems to extract more useful signal with low SNR (-15dB). With higher SNR, our method seems to perform better when discontinuities are rare (**R** and **HS**) but has more difficulties when signals contain frequent discontinuities (**PR**, **PP** and **B**) but remains competitive with [9] with an initial SNR of -5 or 5dB. These results confirm that, despite their simplicity, our synthetic data allow good performances in various problems.

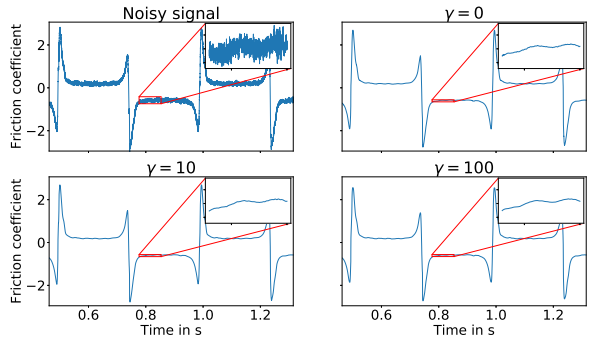
Additionally, inference for these 1s signals is about 100 times faster with our method. While this is not critical for denoising 1024 samples, it would become a huge issue for longer files. A test with a signal of 16384 samples resulted in a 2000 times faster computing with our method.

Lastly, the method of [9] requires the standard deviation of the noise as information to perform denoising, while our method does not need any additional information during inference (as long as the SNR is not too far from the ones of the signals used during training).

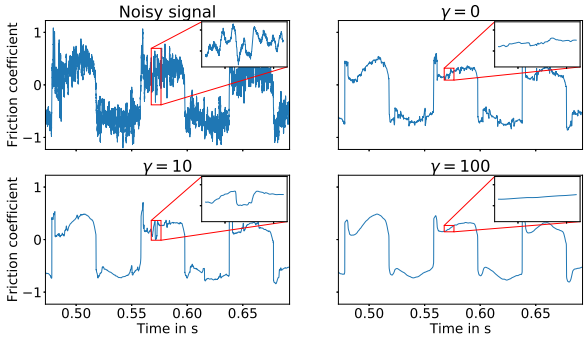
D. Results on real data

To check the generalization of our model, we tested it on friction signals obtained by Ireis (HEF Group)[2] with a linear reciprocating tribometer. Signals were centered then normalized before performing the denoising, then brought back to their original space. We picked the "Medium" model based on the results displayed in Table III.

[2]http://www.ireis.fr/en/



(a) **At low frequency**



(b) **At high frequency**

Fig. 4: Results with real friction data

For these data, ground truth is not available. Therefore, only visual comments are possible. With signals at low frequency (see Figure 4(*a*)), the entire signal, including the peaks, seems to be well denoised. At such frequency, the regularization does not seem to have a huge influence. With signals at high frequency (see Figure 4(*b*)), the influence of regularization is much easier to notice. While large noise remains in the signal denoised without regularization, this pollution is much lower with $\gamma = 10$ and the result is very smooth with $\gamma = 100$. This could be explained by the fact that the real noise is not white, resulting in local peaks, as shown by the zoom on the noisy signal. This result seems to indicate that regularization helps the model generalize the denoising to more realistic noises than the ones used for training.

E. Application to synthetic colored noises

To confirm observations made in previous section, we perform a new evaluation on a set of synthetic signals. The ground truth is generated as in (3) with $f \in \{4, 8, 12\}$Hz. A white noise is first created to reach a predefined SNR randomly picked from (-2.5,0,2.5 or 5dB) and is filtered by applying a mask to its Fourier transform. The low-pass (resp. high-pass) mask takes the value 0 (resp. 4) at 0Hz and 4 (resp. 0)

TABLE VI: SNR obtained with colored noises

	Noisy	$\gamma = 0$	$\gamma = 1$	$\gamma = 10$	$\gamma = 100$
High-pass	-3.84	31.32	29.79	32.16	26.91
Low-pass	-3.81	13.83	13.61	14.44	16.40
Band-pass	-3.92	31.06	31.05	31.20	26.69

at 50000Hz, with a quadratic evolution. The band-pass mask takes the value 0 at 0Hz and 50000Hz and 4 at 25000Hz, with a quadratic evolution on both segments. The resulted noise is added to the ground truth to form the noisy signal. The average SNR of 3000 evaluation signals is around -3.8dB in each set.

We show a comparison of the results obtained with the "medium" architecture and several values of γ in Table VI.

These results show a different behaviour according to the frequencies of noise. Stronger regularization seems to be very helpful with low frequencies, which might be explained by the fact that the model cannot consider a phenomenon with a period longer than its receptive field as a noise. With high frequencies, all models perform well, which could be expected as such frequencies would already be removed efficiently by a moving average. Here, the band-pass mask allows frequencies around 25kHz, which are also high, explaining the good results.

Also notice that the previous observations were less marked with masks taking values from 0 to 1 or 2 (especially with low frequencies) or higher SNR, showing that regularization is more helpful with large noises and when the imbalance between low and high frequencies is higher.

V. Conclusion

In this paper we proposed a WaveNet based deep learning model for denoising periodic signal with discontinuities. This model is well adapted for denoising friction signals obtained through a rotating tribological system. It is trained with synthetic data and includes a regularization term to better generalize to real-world data. Further investigation of this parameter might help find the best compromise between cancellation of noise and preservation of peaks. Additionally, while using periodic data, periodicity has limited impact on the model (the receptive field is not large enough to observe several periods at the same time) and only the discontinuous aspect has a significant impact on the training. Better consideration of the periodicity might be beneficial for the model.

References

[1] X. Wang, Y. Zi, and Z. He, "Multiwavelet denoising with improved neighboring coefficients for application on rolling bearing fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 25, no. 1, pp. 285–304, 2011.

[2] S. Abbasion, A. Rafsanjani, A. Farshidianfar, and N. Irani, "Rolling element bearings multi-fault classification based on the wavelet denoising and support vector machine," *Mechanical Systems and Signal Processing*, vol. 21, no. 7, pp. 2933–2945, 2007.

[3] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[4] C. Li, L. Zhan, and L. Shen, "Friction signal denoising using complete ensemble emd with adaptive noise and mutual information," *Entropy*, vol. 17, no. 9, pp. 5965–5979, 2015.

[5] J. Chen, J. Benesty, Y. A. Huang, and E. J. Diethorn, "Fundamentals of noise reduction," in *Springer Handbook of Speech Processing*, pp. 843–872. Springer, 2008.

[6] L. Condat, "A direct algorithm for 1-D total variation denoising," *IEEE SPL*, vol. 20, no. 11, pp. 1054–1057, 2013.

[7] I. W. Selesnick, A. Parekh, and I. Bayram, "Convex 1-D total variation denoising with non-convex regularization," *IEEE SPL*, vol. 22, no. 2, pp. 141–144, 2014.

[8] I. Selesnick, "Total variation denoising via the moreau envelope," *IEEE SPL*, vol. 24, no. 2, pp. 216–220, 2017.

[9] Y. Ding and I. W. Selesnick, "Artifact-free wavelet denoising: Non-convex sparse regularization, convex optimization," *IEEE SPL*, vol. 22, no. 9, pp. 1364–1368, 2015.

[10] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM TASLP*, vol. 26, no. 9, pp. 1570–1584, 2018.

[11] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 APSIPA ASC*. IEEE, 2017, pp. 006–012.

[12] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5069–5073.

[13] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5024–5028.

[14] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[15] Z. Meng, J. Li, Y. Gong, and B.-H. F. Juang, "Cycle-consistent speech enhancement," *arXiv preprint arXiv:1809.02253*, 2018.

[16] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[17] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM TASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.

[18] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307–1335, 2018.

[19] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 969–977.

[20] A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, and S. Birchfield, "Structured domain randomization: Bridging the reality gap by context-aware synthetic data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7249–7255.

[21] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[22] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[23] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[24] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, "DSLR-quality photos on mobile devices with deep convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3277–3285.

[25] P. Wang, H. Zhang, and V. M. Patel, "SAR image despeckling using a convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1763–1767, 2017.

[26] S. Andersson, A. Söderberg, and S. Björklund, "Friction models for sliding dry, boundary and mixed lubricated contacts," *Tribology international*, vol. 40, no. 4, pp. 580–587, 2007.