INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Explaining deep learning models for speech enhancement

Sunit Sivasankaran[1*], *Emmanuel Vincent*[2] , *Dominique Fohr*[2]

[1] Microsoft, One Microsoft Way, Redmond, WA, USA
[2]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France.
sunit.sivasankaran@microsoft.com, evincent@inria.fr, dfohr@loria.fr

Abstract

We consider the problem of explaining the robustness of neural networks used to compute time-frequency masks for speech enhancement to mismatched noise conditions. We employ the Deep SHapley Additive exPlanations (DeepSHAP) feature attribution method to quantify the contribution of every time-frequency bin in the input noisy speech signal to every time-frequency bin in the output time-frequency mask. We define an objective metric — referred to as the speech relevance score — that summarizes the obtained SHAP values and show that it correlates with the enhancement performance, as measured by the word error rate on the CHiME-4 real evaluation dataset. We use the speech relevance score to explain the generalization ability of three speech enhancement models trained using synthetically generated speech-shaped noise, noise from a professional sound effects library, or real CHiME-4 noise. To the best of our knowledge, this is the first study on neural network explainability in the context of speech enhancement.

Index Terms: Deep learning, speech enhancement, feature attribution, explainable AI

1. Introduction

Speech enhancement models are often trained in a supervised fashion using simulated data. The simulated data is generated by mixing speech and noise at different signal-to-noise ratios (SNRs) and a deep neural network (DNN) is trained to estimate either the speech and noise spectra or a time-frequency mask [1–3]. Different kinds of noises have been used to simulate noisy speech [4–7]. In the CHIME-4 challenge [8] for example, real noises recorded using 6 different microphones were used to simulate noisy speech. A commercially available sound effects library was used in [9]. Synthetically generated noise such as white or pink noise is also often used [10]. This raises the question of what kind of noise is best suited to train the network. Real noise matching the conditions in which the speech enhancement model is to be deployed is a good choice. Yet recording noise scenes that cover all these conditions is expensive and often infeasible. An alternative is to use synthetically generated noise, provided that the impact on the enhancement performance is not drastic.

In this article we show that a speech enhancement model trained with synthetically generated speech-shaped noise (SSN) greatly improves the automatic speech recognition (ASR) performance on the CHiME-4 dataset. We focus on explaining this result, as a first step towards predicting the generalization ability of speech enhancement models and choosing optimal training noises in the future. To do so, we use a *feature attribution* method [11–13] and propose a metric to quantify the importance of each input time-frequency bin in the estimation of the output mask. There are multiple *feature attribution* methods proposed in the literature [14–20]. Many of these methods, such as deconvolution networks [16] and grad-cam [21, 22], are designed for a particular DNN architecture, such as a convolutional neural network (CNN). Others such as DeepLIFT [20] are designed for a wider range of architectures. DeepSHAP [23] combines ideas from SHapley Additive exPlanations (SHAP) [23] and DeepLIFT. It was shown in [23] that multiple *feature attribution* techniques such as layer wise relevance propagation (LRP), local interpretable model-agnostic explanations (LIME) [24] and DeepLIFT are special cases of DeepSHAP, therefore we use DeepSHAP to explain the performance of speech enhancement models in this work. Existing studies on the explanation of neural network models for speech processing have focused on classification tasks [25–28]. To the best of our knowledge, this is the first study on feature attribution for speech enhancement — a sequence-to-sequence regression task.

The rest of the article is organized as follows. Section 2 provides an overview of DeepSHAP. Section 3 describes the application of DeepSHAP to speech enhancement models. Section 4 proposes an objective measure to summarize the obtained feature attribution values. Section 5 details our experimental setup and Section 6 discusses the obtained results. We conclude in Section 7.

2. DeepSHAP

DeepSHAP [23] combines the principles of DeepLIFT and SHAP. Consider an input feature vector $\mathbf{x}$ defined as $\mathbf{x} = [x_1, \ldots, x_D]$, where D is the feature dimension. For easier explanation, we assume that the output of the model $\mathcal{F}(\mathbf{x})$ is a scalar.

SHAP computes the relevance of a particular feature x_d by observing the change in the output with respect to the presence vs. absence of that feature. To avoid retraining the network for every combination of present vs. absent features, the absence of a feature is approximated by replacing it by its expected value. Let $\mathbf{x}' = \{x'_1, \ldots . x'_D\}$ denote the simplified feature vector of dimension D where $x'_d \in \{0, 1\}$ denotes the presence or absence of the corresponding feature in $\mathbf{x}$, and let $h_{\mathbf{x}}(\mathbf{x}') = \mathbf{x}$ denote the mapping function which converts the binary vector $\mathbf{x}'$ to the original space:

$$[h_{\mathbf{x}}(\mathbf{x}')]_d = \begin{cases} x_d & \text{if } x'_d = 1 \\ \mathbb{E}(x_d) & \text{if } x'_d = 0. \end{cases} \quad (1)$$

SHAP approximates the network output as a linear combination of the simplified inputs, $\mathcal{F}(h_{\mathbf{x}}(\mathbf{x}')) \approx \phi_0 + \sum_{d=1}^{D} \phi_d \, x'_d$. Each weight ϕ_d is referred to as a SHAP value. It can be either positive or negative and it directly quantifies the relevance of the corresponding feature.

In practice, the replacement of every absent feature by its expected value is a poor approximation when applied to

[*] work done while being a PhD student at Inria-Nancy

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-1764

the whole network. DeepSHAP combines efficient, analytical computation of SHAP values for simple network modules (linear, maxout, activation) with DeepLIFT's mutiplier composition rule to backpropagate these attribution values down to the input layer.

3. Computing SHAP values for speech enhancement models

In the context of multichannel speech enhancement, the input signal $\mathbf{x}(t)$ is a vector consisting of the signals acquired at I microphones, i.e., $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$ with t the time index, and it contains a single speech source $\mathbf{c}(t)$ and noise $\mathbf{u}(t)$:

$$\mathbf{x}(t) = \mathbf{c}(t) + \mathbf{u}(t). \tag{2}$$

The signals are represented in the time-frequency domain via the short time Fourier transform (STFT). A DNN $\mathcal{F}(\cdot)$ is trained to estimate the ideal ratio mask (IRM)

$$\mathrm{M}(n,f) = \frac{|c_1(n,f)|}{|c_1(n,f)| + |u_1(n,f)|} \tag{3}$$

using a single-channel input magnitude STFT — say channel 1, i.e., $\widehat{\mathbf{M}} = \mathcal{F}(|\mathbf{X}_1|)$, where $|\mathbf{X}_1|$ and $\widehat{\mathbf{M}}$ are $N \times F$ matrices containing the magnitude STFT coefficients of $x_1(t)$ and the estimated mask values $\widehat{\mathrm{M}}(n,f)$ for all time frames $n \in \{1, \ldots, N\}$ and frequency bins $f \in \{1, \ldots, F\}$.

A natural way of using DeepSHAP is to assume that each magnitude STFT coefficient $|x_1(n', f')|$ is an input feature and to compute the contribution of that feature to every time-frequency bin $\widehat{\mathrm{M}}(n,f)$ of the output mask. This results in $N \times F$ relevance matrices $\Phi^{\mathrm{TF}}(n,f)$ of size $N \times F$ each, i.e., one matrix per output time-frequency bin, which we refer to as time-frequency SHAP. In order to reduce the number of matrices to be computed and analyzed, an alternative is to sum up the attribution values across (output) frequency as

$$\Phi^{\mathrm{T}}(n) = \sum_f \Phi^{\mathrm{TF}}(n,f). \tag{4}$$

This doesn't require the computation of every $\Phi^{\mathrm{TF}}(n,f)$. Instead, the SHAP values are summed at the output layer, and a single backpropagation to the inputs is performed. The F resulting matrices $\Phi^{\mathrm{T}}(n)$ — which we refer to as time SHAP — are also $N \times F$ matrices, showing the relevance for every time frame of the output mask. Similarly, attribution values can also be summed over the whole utterance as $\Phi^{\mathrm{U}} = \sum_n \Phi^{\mathrm{T}}(n)$. We refer to Φ^{U} as utterance SHAP. It can be observed that $\sum_{n,f} \Phi^{\mathrm{TF}}(n,f) = \sum_n \Phi^{\mathrm{T}}(n) = \Phi^{\mathrm{U}}$. $\Phi^{\mathrm{TF}}(n,f)$ gives the highest possible granularity of relevance while Φ^{U} gives the lowest possible granularity.

In our preliminary experiments, we have observed that the time-frequency SHAP maps $\Phi^{\mathrm{TF}}(n,f)$ obtained for different frequency bins f in a single frame n are similar to each other, while the utterance SHAP maps Φ^{U} tend to suppress details that were found to be locally relevant. This phenomenon can be observed in Fig. 1. We therefore choose $\Phi^{\mathrm{T}}(n)$ for our following analysis. Fig. 1 also shows that both positive and negative SHAP values point to time-frequency bins dominated by speech. Therefore, we use the absolute value of the SHAP attributions in the rest of this work.

4. Speech relevance score

Analyzing feature attributions is non-trivial. This is usually done by human visualization which is subjective by nature. In this section we propose an objective measure to summarize the SHAP values obtained for speech enhancement models.

The job of a speech enhancement model is to remove the time-frequency bins associated with noise while retaining the time-frequency bins associated with speech. We argue that a well-trained speech enhancement model, with good generalization ability, should mostly look at the time-frequency bins belonging to speech. This is particularly true while evaluating the model in unseen noise conditions, where a speech enhancement model will only have access to speech patterns learned from the training dataset with no prior knowledge about the noise spectra.

Based on this assumption we propose the following measure, which we refer to as the *speech relevance score* (η), to summarize the estimated SHAP values:

$$\eta = \frac{\sum_{n \in \text{speech}} \#\{\phi_{>T+\text{IBM}}(n)\}}{\sum_{n \in \text{speech}} \#\{\phi_{>T}(n)\}}, \tag{5}$$

where $\#\{\phi_{>T}(n)\}$ represents the number of time-frequency bins in $\Phi^{\mathrm{T}}(n)$ whose absolute value is greater than a threshold T and $\#\{\phi_{>T+\text{IBM}}(n)\}$ represents the number of bins among those which are also identified as speech in the ideal binary mask (IBM) defined as

$$\mathrm{IBM}(n,f) = \begin{cases} 1 & \text{if} \quad |c_1(n,f)| > |u_1(n,f)| \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

A visual representation of $\#\{\phi_{>T}(n)\}$ and $\#\{\phi_{>T+\text{IBM}}(n)\}$ for a small speech segment is shown in Fig. 2.

We chose the IBM over the IRM to define η, since η is then akin to the classical *precision* measure used for classification tasks, which is easy to interpret. The measure is computed only for frames containing speech. The threshold T denotes the T-th percentile of the absolute SHAP values in each frame. The number $\#\{\phi_{>T}(n)\}$ of time-frequency bins with large enough SHAP value increases with decreasing T. Note that we use IBM only to compute the speech relevance score while the model is trained to estimate IRM.

As mentioned above, the speech relevance score is akin to a *precision* measure. The computation of a *recall* measure is not feasible here, since there is no ground truth regarding which time-frequency bins *must* contribute to mask estimation. In particular it cannot be assumed that the mask values in a given time frame must depend on all time-frequency bins belonging to speech (in all time frames).

5. Experimental setup

5.1. Dataset

Experiments are conducted using the CHiME-4 dataset [8], which consists of Wall Street Journal sentences spoken by talkers situated in challenging noisy environments recorded using a 6-channel tablet-based microphone array. The original dataset considers four different categories of environments: bus, cafe, pedestrian area, and street junction. It comes with a data simulation tool, which mixes original non-reverberated WSJ0 utterances with background noise, ensuring the same SNR distribution as real noisy recordings on every channel.

In order to train the speech enhancement network, we generate three different training datasets corresponding to three different noise conditions, namely

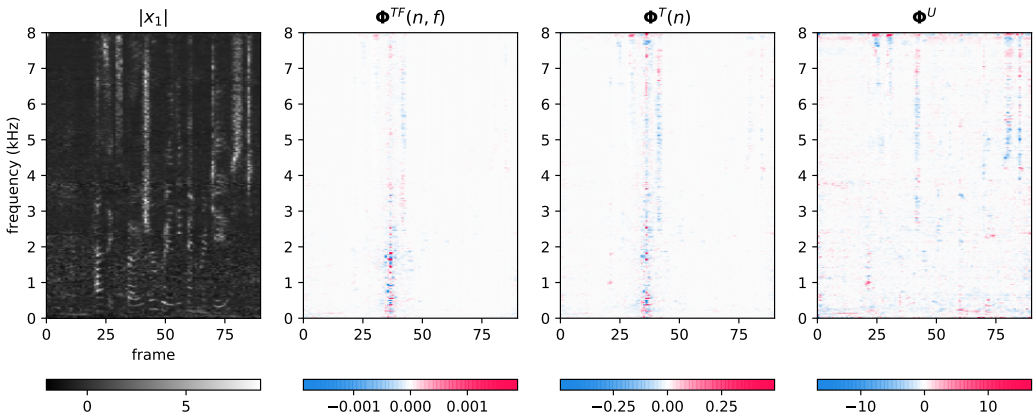

Figure 1: *Example SHAP values computed for a noisy speech signal. The four subplots (from left to right) represent the input spectrogram after mean and variance normalization, the time-frequency SHAP map $\Phi^{TF}(n, f)$ for $n = 36$ and $f = 1635$ Hz, the time SHAP map $\Phi^{T}(n)$ for $n = 36$ and the utterance SHAP map Φ^{U}, respectively.*

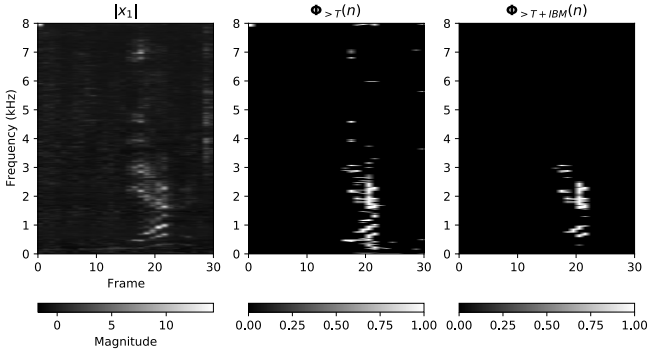

Figure 2: *Input spectrogram, $\phi_{>T}(n)$ and $\phi_{>T+IBM}(n)$ used to compute speech relevance score for a speech signal containing 30 frames.*

1. CHiME: real CHiME-4 noise recordings
2. Speech shaped noise (SSN): a form of synthetically generated noise which is obtained by applying a "speech-like" filter to white noise [29–31]. In the following, a different speech-like filter is computed for every noisy utterance by randomly drawing 6 clean speech utterances and averaging their STFT magnitude spectra across all time frames
3. Network: noise from the Network Sound Effects library[1] as used in [9], containing sounds from various categories such as music, weather, rail, etc.

There are $7,138$ utterances for training and $1,640$ for validation for each noise condition. The distribution of SNR values is the same across all conditions and as in the original CHiME-4 dataset.

5.2. DNN architectures for speech enhancement

The DNNs for speech enhancement are trained to estimate the IRM using the STFT magnitude spectra of x_5 (i.e., channel 5 of the multichannel mixture signal) as inputs. The STFT window size was 50 ms with a 25 ms shift. The input dimension of the network was 401. The input STFT magnitude spectra was mean and variance normalized.

The DNN architecture contained two bidirectional long short term memory (Bi-LSTM) layers followed by layer normalization [32] and a linear layer containing 401 hidden units. The output was constrained to lie in the range $[0 - 1]$ using a sigmoid nonlinearity. The speech enhancement models trained on the three different noise conditions are denoted as $\mathcal{F}_{\text{CHIME}}$, $\mathcal{F}_{\text{SSN}}$, and $\mathcal{F}_{\text{NETWORK}}$, respectively. The estimated mask along with multichannel signals from all channels (with the exception of channel 2) is used to compute a rank-1 constrained multichannel Wiener filter [33] which is then used to obtain the enhanced speech.

5.3. ASR evaluation

We evaluate the ASR performance resulting from speech enhancement on the real evaluation set (`et05_real`) and the simulated development set (`dt05_simu`) of the original CHiME-4 dataset using the baseline ASR system provided by the CHiME-4 challenge organizers. This system follows the `nnet1` recipe of the Kaldi ASR toolkit [34], involving a 7-layer multi layered perceptron (MLP)-based acoustic model and a 3-gram language model. It was trained on both real and simulated noisy speech.

5.4. Speech relevance score computation

We compute the speech relevance score on several simulated datasets for which the ground truth IBM is known, including the original CHiME-4 simulated evaluation set (`dt05_simu`). The SHAP toolkit[2] was used to compute SHAP values. The expected value of the inputs at every layer was computed over 40 random utterances from the considered dataset. For example, if SHAP values are to be computed for an utterance in `dt05_simu`, then 40 random utterances from `dt05_simu` are used for that purpose. We used the DeepExplainer[3] component of the toolkit, which computes the SHAP values analytically for simple DNN modules (linear, sigmoid) or using the gradient for complex modules (Bi-LSTM) and backpropagates them with DeepLIFT's multiplier composition rule. A threshold value of $T = 99.9$ is used unless mentioned otherwise (see Section 6.2).

[1] `https://www.sound-ideas.com/Product/199/Network-Sound-Effects-Library`

[2] `https://github.com/slundberg/shap`

[3] `https://github.com/slundberg/shap/blob/master/shap/explainers/deep/deep_pytorch.py`

A high value of T results in a lower number of time-frequency bins to compute η.

6. Results

6.1. ASR

Table 1 shows the ASR results on the `et05_real` dataset using different speech enhancement models. The same ASR model was used for all the experiments. A baseline WER of 25.9% was obtained when no speech enhancement was performed. The WER improved to 11.7% by enhancing speech using the $\mathcal{F}_{\text{CHIME}}$ model. The WERs obtained using the speech enhancement models trained with SSN and Network noise are 14.0% and 15.1%, respectively. The improved performance with the speech enhancement model trained using CHiME noise can be attributed to the matched condition between the training and evaluation data. Nevertheless, the results obtained using speech enhancement models trained with SSN and Network noise are significantly better than the baseline showing the usefulness of these noises for training a speech enhancement model. Similar gains in the ASR performance can be observed on the `dt05_simu` dataset.

Table 1: *WER (%) on the CHiME-4 real evaluation (`et05_real`) and simulated development (`dt05_simu`) datasets.*

Training Noise	`et05_real` (%)	`dt05_simu`(%)
Baseline (No Enhancement)	25.9	12.7
CHiME	11.7	6.7
SSN	14.0	7.3
Network	15.1	7.7

6.2. Speech relevance score

Table 2 shows the speech relevance score obtained on `dt05_simu` (η cannot be computed for `et05_real` due to non-availability of the corresponding clean speech) for all the speech enhancement models with different threshold values. The results are obtained using a total of 300 utterances. A speech relevance score of 94.8% was obtained using $\mathcal{F}_{\text{CHIME}}$, meaning that for a threshold $T = 99.9$, 94.8% of the time-frequency bins in the input spectrogram which were used to explain the output mask were dominated by speech. The speech relevance score values in Table 2 follow the trends observed in the ASR results of Table 1. Better η values are seen for the $\mathcal{F}_{\text{CHIME}}$ model, which gave the best ASR performance. The negligible difference between the speech relevance score values for $\mathcal{F}_{\text{SSN}}$ and $\mathcal{F}_{\text{NETWORK}}$ reflects the difference in the ASR results of the corresponding models on `dt05_simu`, albeit in favor of $\mathcal{F}_{\text{NETWORK}}$. The speech relevance score varies with respect to the threshold T, indicating that the time-frequency bins with lower SHAP values are not dominated by speech. We can therefore conclude that $\mathcal{F}_{\text{CHIME}}$ works better than other models because it relies on speech-dominated time-frequency bins to estimate the mask.

6.3. Generalization capability of speech enhancement models

To better understand how the models generalize, we create two different simulated datasets which we refer to as *Gen-train* and *Gen-test* set. The *Gen-train* set contains noisy speech obtained by mixing clean speech signal with noise from the same type as the one used to train the model (matched noise). The noisy speech in the *Gen-test* set, on the other hand, are obtained by mixing the clean speech signal with CHiME noise (unmatched noise). SHAP values are computed for $\mathcal{F}_{\text{SSN}}$ and $\mathcal{F}_{\text{NETWORK}}$ models in both the setups. A model with good generalization ability will have a lower difference in the speech relevance score between the *Gen-train* and *Gen-test*.

Table 2: *Speech relevance score (η) (%) values using different thresholds on `dt05_simu`.*

Model	$T = 99.9$	$T = 99.0$	$T = 98.0$
$\mathcal{F}_{\text{CHIME}}$	94.8	92.2	90.5
$\mathcal{F}_{\text{SSN}}$	89.6	87.2	85.4
$\mathcal{F}_{\text{NETWORK}}$	90.3	89.5	88.7

Table 3: *Average speech relevance score obtained on* 300 *random utterances in the training and test setups. The speech relevance score for $\mathcal{F}_{\text{CHIME}}$ was 80.5%.*

Experiment setup	$\mathcal{F}_{\text{NETWORK}}$ (%)	$\mathcal{F}_{\text{SSN}}$ (%)
Gen-Train	81.7	86.2
Gen-Test	68.3	77.0
Change (%)	16.4	10.7

Table 3 shows the average speech relevance score computed on *Gen-Train* and *Gen-Test* datasets containing 300 samples each. $\mathcal{F}_{\text{SSN}}$ has a lower change in speech relevance score (10.7%) as compared to $\mathcal{F}_{\text{NETWORK}}$ (16.4%) indicating better generalization capability for $\mathcal{F}_{\text{SSN}}$.

7. Conclusion

In this article we approached the problem of explaining the predictions of a speech enhancement model. DeepSHAP, a *feature attribution* method, is employed to figure out which time-frequency bins of the input spectrogram are used by the DNN to estimate the mask. Based on the idea that a well-trained model should look at time-frequency bins dominated by speech instead of those dominated by noise, we proposed speech relevance score — a measure to evaluate *feature attributions*. We showed that speech enhancement models having a higher speech relevance score give better ASR performance. We also showed that the generalization capability of a speech enhancement model trained using synthetically generated SSN is better than that of a speech enhancement model trained using Network noise.

8. Acknowledgements

This work was made with the support of the French National Research Agency, in the framework of the project VOCADOM "Robust voice command adapted to the user and to the context for AAL" (ANR-16-CE33-0006). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

9. References

[1] S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, *New Era for Robust Speech Recognition: Exploiting Deep Learning*. Springer, 2017.

[2] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Academic Press, 2015.

[3] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[4] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv:2005.11262 [eess.AS]*, 2020.

[5] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," in *Interspeech*, 2019, pp. 1368–1372.

[6] S. Sivasankaran, E. Vincent, and D. Fohr, "Analyzing the impact of speaker localization errors on speech separation for automatic speech recognition," in *28th European Signal Processing Conference*, 2021.

[7] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The Interspeech 2020 Deep Noise Suppression Challenge: Datasets, subjective testing framework, and challenge results," in *Interspeech*, 2020, pp. 2492–2496.

[8] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, Nov. 2017.

[9] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.

[10] A. Abdulaziz and V. Kepuska, "Noisy TIMIT speech," *Linguistic Data Consortium V1 , http://hdl.handle.net/11272/UFA9N*, 2017.

[11] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[12] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020.

[13] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, 2020, https://distill.pub/2020/attribution-baselines.

[14] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *International Conference on Learning Representations (Workshop Poster)*, 2014.

[15] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," in *ICML Workshop on Visualization for Deep Learning*, 2017.

[16] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.

[17] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *International Conference on Learning Representations (Workshop Track)*, 2015.

[18] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, 2017, pp. 3319–3328.

[19] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1932–6203, Jul. 2015.

[20] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, 2017, pp. 3145–3153.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[22] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 839–847.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[25] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple DoA estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, Mar. 2019.

[26] H. Muckenhirn, V. Abrol, M. Magimai.-Doss, and S. Marcel, "Understanding and visualizing raw waveform-based CNNs," in *Interspeech*, 2019, pp. 2345–2349.

[27] H. Bharadhwaj, "Layer-wise relevance propagation for explainable deep learning based speech recognition," in *IEEE International Symposium on Signal Processing and Information Technology*, 2018, pp. 168–174.

[28] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *arXiv:1807.03418 [cs, eess]*, Jul. 2018.

[29] M. Pariente and D. Pressnitzer, "Predictive denoising of speech in noise using deep neural networks," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2611–2611, Oct. 2017.

[30] F. Li, P. S. Nidadavolu, and H. Hermansky, "A long, deep and wide artificial neural net for robust speech recognition in unknown noise," in *Interspeech*, 2014, pp. 358–362.

[31] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Interspeech*, 2016, pp. 352–356.

[32] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450 [stat.ML]*, 2016.

[33] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Computer Speech & Language*, vol. 49, pp. 37–51, May 2018.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.