FRS: Adaptive Score for Improving Acoustic Source Classification From Noisy Signals

R. Marinati, *Student Member, IEEE*, R. Coelho, *Senior Member, IEEE*, and L. Zão, *Member, IEEE*

***Abstract*—This letter introduces a Frame Relevance Score (FRS) to improve the classification of environmental acoustic sources from noisy speech signals. The importance of each short-time frame for the classification results is objectively interpreted by SHapley Additive exPlanations (SHAP) values. The FRS enables the selection of frames that are more appropriate to improve the discrimination power of the acoustic models. The FRS-based frame selection can be used as a pre-training strategy to any classification approach. Evaluation experiments consider the recognition of ten background sources from noisy speech signals. The classical system based on MFCC and GMM is adopted to prove that the selected frames can better distinguish the acoustic classes. Moreover, the proposed solution outperforms a surrogate-based adaptive learning technique and a competing frame selection method. Experiments are also conducted with a recently proposed pre-trained neural network that achieves high classification rates. For this scenario, the FRS-based selection improves the overall classification accuracy from 51.5% to 58.8%.**

***Index Terms*—Acoustic source classification, noisy speech signals, surrogates, convolutional neural networks.**

I. Introduction

In the last decade, the classification of acoustic sources and scenes has gained significant attraction [1], [2], [3], [4], [5]. Hearing aid, robot navigation, and smart devices are example applications of this important task. Most of these studies mainly focus on machine learning approaches [3], [5], such as dictionary learning and convolutional neural networks (CNN). However, large amount of data may be required for the training. Thus, the analysis about the relevance of the observations available to generate the source models is a key challenge.

The recognition of acoustic sources is particularly important for speech-processing systems when speech signals are captured in real noisy scenarios [6]. Several research works show that speech enhancement [7], [8], [9], [10], [11], [12], speech recognition [13], and speaker identification [14], [15] are largely affected by acoustic background noises. Their results significantly vary according to the characteristics of the corrupting noise. Thus, the classification of the acoustic source that corrupts the speech signal may improve robustness of speech-based applications.

This Letter proposes a solution to improve the accuracy of acoustic sources classification from noisy speech signals. For this purpose, an objective score is introduced based on SHapley Additive exPlanations (SHAP) [16] values to define the most relevant frames of a target signal. The first step is to train a CNN using the original unprocessed training data. Matrices of SHAP values are then computed from each training signal to assign the importance of each input feature to the network output. In this work, the Frame Relevance Score (FRS) is defined according to the SHAP values computed from the target and the remaining classes. This objective score enables the removal of the least relevant frames from the training dataset to achieve better discrimination power among classes. The FRS-based frame selection is suitable as a pre-training solution for any classification approach.

Evaluation experiments are conducted in two scenarios according to the input signals: acoustic sources only and noisy speech signals. For both cases, two different approaches are applied to classify ten acoustic sources. The first system is based on mel-frequency cepstral coefficients (MFCC) [17] and Gaussian mixture models (GMM), which is an important stochastic strategy for a variety of acoustic classification applications [14], [18], [19]. The second approach is the recently introduced Pretrained Audio Neural Network (PANN) [3]. MFCC + GMM results show that the proposed solution improves the classification accuracy in both scenarios. For the classification from noisy speech signals, the FRS-based frame selection also outperforms two pre-training methods adopted as baseline. In terms of PANN, classification accuracies are higher than those attained with the classical MFCC + GMM system. PANN results show that the FRS-based frame selection provides substantial improvement for noisy speech signals.

The main contributions of this work are summarized as follows:

- Introduction of an objective Frame Relevance Score for acoustic sources classification based on SHAP values;
- Definition of a threshold to detect the least relevant frames to enable a pre-training solution for any classification approach;
- Evaluation of the proposed frame selection for two different classification approaches: MFCC + GMM and PANN.

II. SHAP-Based Frame Relevance Score

The definition of the most relevant frames for acoustic sources classification is based on the computation of SHAP values [16]. SHAP is a unified framework to interpret complex models

Manuscript received 18 October 2023; revised 22 December 2023; accepted 11 January 2024. Date of publication 24 January 2024; date of current version 1 March 2024. This work was supported in part by the National Council for Scientific and Technological Development (CNPq) under Grant 305488/2022-8, in part by Fundação de Amparo Pesquisa do Estado do Rio de Janeiro (FAPERJ) under Grant 200518/2023, and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) under Grant 001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui. *(Corresponding author: R. Coelho.)*

The authors are with the Laboratory of Acoustic Signal Processing, Military Institute of Engineering, Rio de Janeiro 22290-270, Brazil (e-mail: coelho@ime.eb.br).

Digital Object Identifier 10.1109/LSP.2024.3358097

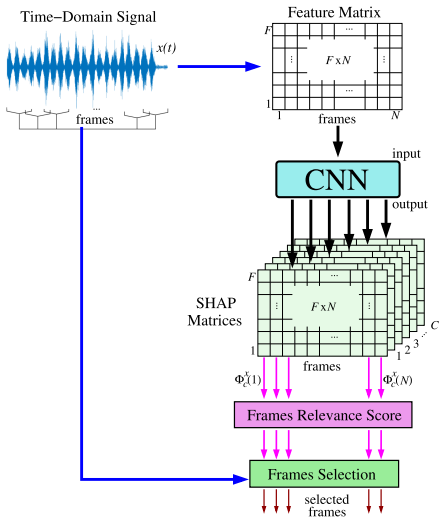

Fig. 1. Block diagram of the FRS-based frames selection.

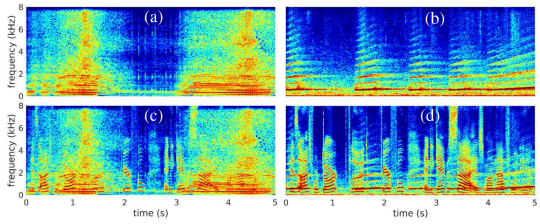

Fig. 2. Spectrograms of (a) chainsaw and (b) siren acoustic sources, and a speech signal corrupted by the same (c) chainsaw and (d) siren sources with SNR of 0 dB.

predictions, as those obtained from deep learning. It generalizes six other feature attribution methods such as LIME [20] and DeepLIFT [21]. In [22], SHAP was adopted to define a speech relevance score to serve as an objective speech enhancement measure.

Fig. 1 depicts the block diagram of the solution introduced in this Letter. Spectrograms in Fig. 2 are used to illustrate the challenging task of defining frames relevance to improve the acoustic sources classification task. Spectrograms in Fig. 2(a), (b), obtained from two acoustic sources (chainsaw and siren), show that such sources present different behavior in terms of time-frequency representation. This behavior does not hold when these sources corrupt a speech signal, as in Fig. 2(c), (d). The noisy speech signals consider a signal-to-noise ratio (SNR) of 0 dB. It may be noted that, for example, the chainsaw source is more prominent in Fig. 2(c) in time frames around $[1.0 - 1.5]$ s and $[4.0 - 5.0]$ s. On the other hand, frames where the siren noise prevail over the speech signal are not easily found by the visual inspection of Fig. 2(d).

A. SHAP Values

SHAP values interpret the importance of a given feature for a particular prediction by observing how its presence or absence affects the network output. The main issue is to explain a prediction $f(\mathbf{x})$ based on a single D-dimensional input feature $\mathbf{x}$. For this purpose, let $\mathbf{x} = h_{\mathbf{x}}(\mathbf{x}')$ denote the mapping function between $\mathbf{x}$ and the simplified input $\mathbf{x}' \in \{0,1\}^D$. In this binary vector $\mathbf{x}'$, the values 0 and 1 respectively denote the absence and presence of the corresponding feature. When a component $\mathbf{x}_d$ of $\mathbf{x}$ is absent, the mapping function $h_{\mathbf{x}}()$ is approximated by its expected value, such that

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

As an additive feature attribution method, SHAP approximates the network output as

$$f(\mathbf{x}) = f\left(h_{\mathbf{x}}(\mathbf{x}')\right) = \phi_0 + \sum_{d=1}^{D} \phi_d \mathbf{x}'_d. \tag{2}$$

The weights $\phi_d \in \mathbb{R}$ in (2) correspond to the SHAP values of the input features $\mathbf{x}_d, d = 1, \ldots, D$.

In [16], different approaches are presented to solve the approximation problem and compute the SHAP values. Particularly, the Deep SHAP method connects Shapley values and DeepLIFT [21] in a high-speed approximation algorithm for deep learning models. Due to this reason, the Deep SHAP implementation from the SHAP toolkit[1] is adopted to compute the SHAP values in this work.

B. Frame Relevance Score (FRS)

Consider a convolutional neural network trained to classify acoustic sources among C different classes (refer to Fig. 1). Let $F \times N$ be the size of the input feature matrices, where F is the number of feature coefficients extracted from each frame, and N is the total number of frames. For each acoustic signal $x(t)$ available for training, SHAP values are computed to form a set of C matrices $\{\Phi^x_c, c = 1, \ldots, C\}$. One $F \times N$ matrix of SHAP values is obtained for each class in the CNN output layer. Let $c_x \in \{1, \ldots, C\}$ denote the index of the class that $x(t)$ belongs to. For each frame $n = 1, \ldots, N$, define $\Phi^x(n)$ as the sum of all SHAP values from the n-th column of $\Phi^x_{c_x}$,

$$\Phi^x(n) = \sum_{f=1}^{F} \Phi^x_{c_x}(f, n), c = 1, \ldots, C. \tag{3}$$

Similarly, let $\Phi^x_{\text{aver}}(n)$ denote the sum of the n-th column values of the SHAP matrices computed from $x(t)$ averaged over the C classes. Thus,

$$\Phi^x_{\text{aver}}(n) = \frac{1}{C} \sum_{c=1}^{C} \sum_{f=1}^{F} \Phi^x_c(f, n). \tag{4}$$

Finally, the Frame Relevance Score is here introduced as

$$\text{FRS}^x(n) = \Phi^x(n) - \Phi^x_{\text{aver}}(n), \tag{5}$$

which is expected to indicate the importance of the n-th frame of $x(t)$ to the correct decision prediction of class c_x.

III. FRS-Based Frame Selection

In this work, the Frame Relevance Score is employed to detect and remove the least informative frames for the training models. The idea is to apply the proposed frame selection only to the acoustic signals available for training. As stated in Section II-B, frames of a training signal $x(t)$ with greater FRS values indicate

[1][Online]. Available: https://github.com/slundberg/shap

increased relevance to the correct prediction of the class it belongs to. Thus, a selection threshold θ must be set such that the n-th frame of $x(t)$ is preserved whenever $\text{FRS}^x(n) > \theta$, and discarded otherwise. For this purpose, let μ_{FRS} and σ_{FRS} denote the mean and standard deviation of the FRS values computed from all frames of every signal available for training. The threshold is here defined as

$$\theta = \mu_{\text{FRS}} - \sigma_{\text{FRS}}\,, \tag{6}$$

which removes those frames whose FRS values are significantly lower than the average.

The proposed FRS-based frame selection is suitable as a pre-training solution for any classification strategy. For classical stochastic approaches, such as GMM, the idea is to remove from the feature matrices those vectors extracted from frames with low FRS values. In this work, the MFCC + GMM approach considers MFCC matrices with fewer columns to train the GMM of each class.

For the classification using the PANN pre-trained network, the input feature matrices must have the same size for both training and test phases. It means that feature vectors from the low FRS frames should not be simply discarded in such classification approaches. To this end, a surrogate sample sequence is generated as in [2], [23] to reproduce the Kurtosis ratio, the power spectral density, and the index of nonstationarity [24] of the target signal. Each training signal is then reconstructed in the time domain considering only the most relevant frames, while the discarded ones are replaced by the corresponding frames from the surrogate sequence. The Hanning window is applied to ensure the continuity of the reconstructed signal.

IV. Experiments and Results

Acoustic sources classification experiments are conducted to evaluate the proposed FRS-based frame selection method. For this purpose, a subset of the ESC dataset [25] composed by 400 audio recordings are selected from ten different classes: airplane (Air), bells (Bel), chainsaw (Cha), engine (Eng), hand saw (Han), helicopter (Hel), siren (Sir), train (Tra), vacuum cleaner (Vac), and washing machine (Was). Note that these ten classes are recorded from single acoustic sources, i.e., audio samples captured in natural scenes are not applied for the classification system. As in [25], the subset is divided into five folds for cross-validation: while one fold is separated for tests, the remaining folds are used for training.

The classification of acoustic sources is also considered for noisy speech scenarios. For these experiments, 12 speech signals (six female, six male) with time duration greater than 5 seconds collected from the TIMIT [26] database are corrupted with the acoustic sources. To this end, ESC audio recordings are downsampled to 16 kHz, while TIMIT speech signals are cut to 5 seconds length. Three SNR values are adopted for the noisy speech signals, -5 dB, 0 dB, and 5 dB, which are only considered for the tests.

The surrogate assisted training (SAT) [2] and a feature vector-based frame selection (FvFS) are adopted as baseline. The SAT is implemented considering a set of 12 surrogate sequences artificially generated according to [2], [23]. For the surrogates selection, acoustic models are obtained for each training signal and corresponding surrogates sequences. Considering the training dataset, a surrogate sequence is selected whenever it leads to a higher classification accuracy than the original signal. Finally, the selected surrogates are used in the place of the original signals for training.

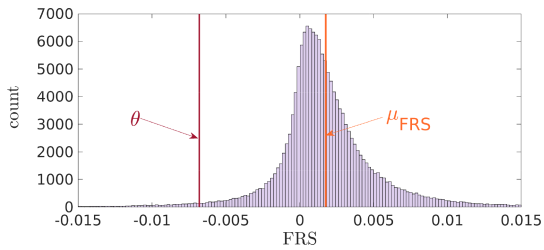

Fig. 3. FRS histogram from acoustic signals available for training the CNN of the first cross-validation experiment.

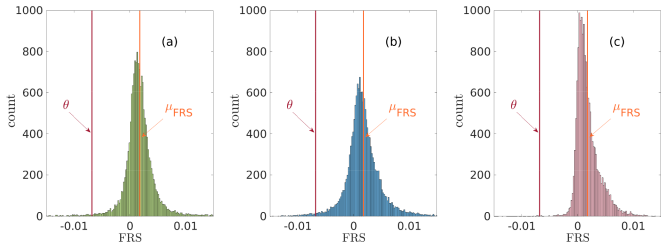

Fig. 4. FRS histograms of three classes: (a) washing machine, (b) chainsaw, and (c) siren.

The FvFS is based on the work introduced in [27]. For the frames selection, four female and four male speech signals from the TIMIT database are corrupted by the acoustic sources considering five SNR values between -10 dB and 10 dB. Let $D(n)$ denote the set of Euclidean distances between the n-th MFCC vector extracted from the acoustic source $x(t)$ and the corresponding vectors from the 40 speech signals corrupted by $x(t)$. The n-th frame of $x(t)$ is selected whenever the sum of distances in $D(n)$ is below some threshold. Ten different threshold values in $[\mu_D, \mu_D + \sigma_D]$ are considered for the FvFS, where μ_D and σ_D denote the mean and standard deviation of the sum of distances computed from all frames of every signal available for training. The FvFS results presented in this work consider the optimal choice of the threshold according to the overall classification accuracy in the noisy speech signal scenario.

A. Analysis of the FRS Values

SHAP values are here computed from the original acoustic sources using the CNN architecture that serves as baseline for the acoustic scene classification task of the DCASE 2018 Challenge [28]. The CNN input size of 40×500 is achieved by extracting 40 log mel-band energies from frames with duration of 40 ms and shift of 10 ms. Thus, ten 40×500 SHAP matrices are obtained for each training signal. The Frames Relevance Scores are then computed according to (3)–(5). Fig. 3 depicts the histogram of the FRS values from the acoustic signals of folds 2–5, that are applied to train the CNN for the first cross-validation experiment. The mean and standard deviation of these FRS values are used to define the threshold θ for the frames selection, also shown in Fig. 3.

In order to illustrate differences in the FRS values among classes, Fig. 4 presents FRS histograms from the washing machine, chainsaw, and siren sources. Note that siren signals have

TABLE I
ACOUSTIC SOURCES CLASSIFICATION ACCURACIES (%) WITH THE CNN ADOPTED TO COMPUTE SHAP VALUES

actual source	classified source									
	Air	Bel	Cha	Eng	Han	Hel	Sir	Tra	Vac	Was
Air	**75.0**	2.5	5.0	0.0	0.0	5.0	0.0	2.5	0.0	10.0
Bel	7.5	**80.0**	0.0	0.0	0.0	0.0	12.5	0.0	0.0	0.0
Cha	2.5	0.0	**55.0**	12.5	0.0	0.0	0.0	2.5	7.5	20.0
Eng	5.0	0.0	7.5	**52.5**	0.0	10.0	0.0	5.0	7.5	12.5
Han	0.0	2.5	0.0	0.0	**92.5**	0.0	0.0	0.0	2.5	2.5
Hel	15.0	0.0	0.0	10.0	0.0	**30.0**	2.5	12.5	0.0	30.0
Sir	0.0	12.5	0.0	0.0	0.0	0.0	**87.5**	0.0	0.0	0.0
Tra	2.5	0.0	0.0	0.0	5.0	20.0	0.0	**70.0**	0.0	2.5
Vac	0.0	0.0	15.0	2.5	0.0	0.0	0.0	0.0	**75.0**	7.5
Was	2.5	0.0	10.0	22.5	2.5	17.5	0.0	5.0	5.0	**35.0**

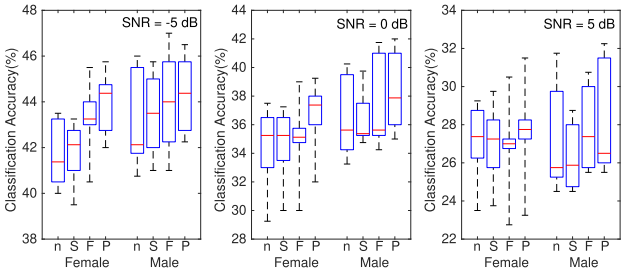

Fig. 5. Classification results obtained from noisy speech signals by MFCC + GMM with (n) no pre-training, (S) SAT, (F) FvFS, and (P) proposed FRS-based frame selection.

fewer frames below the threshold when compared to the other classes. It means that the FRS-based selection criterion discards 0.2% of the frames for siren, while 1.6% and 2.2% of the frames are discarded for chainsaw and washing machine, respectively. These results are consistent with the classification accuracies in Table I: 87.5% of siren signals are correctly recognized, against 55.0% of chainsaw and 35.0% of washing machine. The average accuracy is 65.25%.

B. Pre-Training for MFCC + GMM Classification

The classical MFCC + GMM approach is adopted for the first set of acoustic sources classification experiments from noisy speech signals. For this classification system, feature vectors composed by 20 MFCC are extracted every 10 ms from frames of 40 ms. Each GMM is trained with 8 Gaussian densities considering diagonal covariance matrices.

The accuracy obtained for sources classification from noisy speech signals are presented in Fig. 5. Note that the FRS-based solution achieves the best results for most of the noisy conditions. The overall accuracy is improved from 35.0% to 36.5% when compared to the MFCC + GMM approach with no pre-training. These overall results are computed from 14400 tests, i.e., 12 female and male speech signals corrupted by 400 acoustic sources considering three SNR values. It corresponds to an accuracy precision of 0.019 using the Chebyshev inequality for a confidence degree of 95%. Furthermore, average gains of 0.8 and 1.6 percentage points (p.p.) are achieved over the baseline FvFS and SAT, respectively.

Considering the scenario with acoustic sources only, the classification accuracy is improved from 45.5% with the classical MFCC + GMM, to 46.5% with the baseline FvFS and the proposed FRS frame selection. It is worth to mention that, although the SAT does not improve the overall accuracy from the noisy speech signals, it leads to the best result in the acoustic sources only scenario: 47.0%.

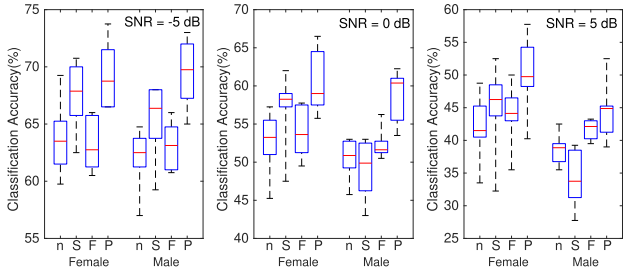

Fig. 6. Classification results obtained from noisy speech signals by the 14-layer PANN with (n) no pre-training, (S) SAT, (F) FvFS, and (P) proposed FRS-based frame selection.

C. Pre-Training for PANN Classification

Acoustic sources classification experiments are also conducted considering the pre-trained 14-layer CNN (PANN) provided by [3]. To this end, the output fully-connected layer of the PANN is adapted to consider the actual number of classes (ten). The CNN is finetuned considering random initial weights for this last layer, while all other parameters are initialized from the pre-trained network. The same set of initial weights are adopted for all experiments. In terms of acoustic sources only scenario, the PANN achieves an classification accuracy of 82.25%. This result is improved to 85.25% with the proposed FRS-based frame selection, against 83.75% with the baseline FvFS. The baseline SAT leads to the best result for this scenario: 86.00%.

The classification results from noisy speech signals obtained with PANN, baseline, and proposed pre-training methods are presented in Fig. 6. The FRS-based frame selection leads to the best result for female and male speech signals and all three SNR values. When compared to the PANN with no pre-training, the average improvement achieves 8.2 p.p. for female speech with SNR of 5 dB, and 7.5 p.p. for male speech with SNR of −5 dB. In terms of overall accuracy, the proposed pre-training method achieves a significant gain of 7.3 p.p., from 51.5% to 58.8%. This result is 5.8 p.p. greater than those obtained by both baseline pre-training solutions.

V. CONCLUSION

This letter proposed the adoption of SHAP values to define a Frame Relevance Score in the context of acoustic sources classification. The acoustic models training is performed only with the most relevant frames to improve the discrimination power among classes. The proposed solution was evaluated in classification experiments divided into two scenarios, according to the input signals: acoustic sources only, and noisy speech signals. The identification of the environmental acoustic source from noisy speech signals would lead to advances in many speech-based applications. Experiments using the classical MFCC + GMM stochastic approach showed that the FRS-based frame selection leads to an overall improvement of 1.5 p.p. in the classification accuracy. The FRS-based frame selection also showed improved results for the Pretrained Audio Neural Network, for which the average accuracy was enhanced from 51.5% to 58.8%.

REFERENCES

[1] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.

[2] G. Zucatelli, R. Coelho, and L. Zão, "Adaptive learning with surrogate assisted training models for acoustic source classification," *IEEE Sens. Lett.*, vol. 3, no. 6, Jun. 2019, Art. no. 7000804.

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[4] G. Zucatelli and R. Coelho, "Adaptive learning with surrogate assisted training models using limited labeled acoustic sample sequences," in *Proc. IEEE Stat. Signal Process. Workshop*, 2021, pp. 21–25.

[5] B. Bahmei, E. Birmingham, and S. Arzanpour, "CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 29, pp. 682–686, 2022.

[6] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame level noise classification in mobile environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 237–240.

[7] A. Caldeira and R. Coelho, "EEMD-IF based method for underwater noisy acoustic signals enhancement in time-domain," *IEEE Signal Process. Lett.*, vol. 30, pp. 294–298, 2023.

[8] C. Medina, R. Coelho, and L. Zão, "Impulsive noise detection for speech enhancement in HHT domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2244–2253, 2021.

[9] R. Tavares and R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 6–10, Jan. 2016.

[10] L. Zão, R. Coelho, and P. Flandrin, "Speech enhancement with EMD and hurst-based mode selection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 5, pp. 899–911, May 2014.

[11] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[12] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[13] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1987, pp. 705–708.

[14] L. Zão and R. Coelho, "Colored noise based multicondition training technique for robust speaker identification," *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 675–678, Nov. 2011.

[15] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.

[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[18] A. Venturini, L. Zão, and R. Coelho, "On speech features fusion, α-integration gaussian modeling and multi-style training for noise robust speaker classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1951–1964, Dec. 2014.

[19] L. Zão, D. Cavalcante, and R. Coelho, "Time-frequency feature and AMS-GMM mask for acoustic emotion classification," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 620–624, May 2014.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[21] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[22] S. Sivasankaran, E. Vincent, and D. Fohr, "Explaining deep learning models for speech enhancement," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 676–679.

[23] L. Zão and R. Coelho, "Generation of coloured acoustic noise samples with non-Gaussian distribution," *IET Signal Process.*, vol. 6, no. 7, pp. 684–688, Sep. 2012.

[24] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3459–3470, Jul. 2010.

[25] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd Annu. ACM Conf. Multimedia*, 2015, pp. 1015–1018.

[26] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop Speech Recognit.*, 1986, pp. 93–99.

[27] S. Song, S. Zhang, B. W. Schuller, L. Shen, and M. Valstar, "Noise invariant frame selection: A simple method to address the background noise problem for text-independent speaker verification," in *Proc. Int. Joint Conf. Neural Netw.*, 2018, pp. 1–8.

[28] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2018, pp. 9–13.