electronics

MDPI

Article

Speech Enhancement Based on Enhanced Empirical Wavelet Transform and Teager Energy Operator

Piotr Kuwałek [1,*] and Waldemar Jęśko [2,3]

[1] Institute of Electrical Engineering and Electronics, Poznan University of Technology, 60-965 Poznan, Poland
[2] Institute of Computing Science, Poznan University of Technology, 60-965 Poznan, Poland; waldemar.jesko@psnc.pl
[3] Poznan Supercomputing and Networking Center, 61-139 Poznan, Poland
* Correspondence: piotr.kuwalek@put.poznan.pl

Abstract: This paper presents a new speech-enhancement approach based on an enhanced empirical wavelet transform, considering the time and scale adaptation of thresholds for individual component signals obtained from the used transform. The time adaptation is performed using the Teager energy operator on the individual component signals, and the scale adaptation of thresholds is performed by the modified level-dependent threshold principle for the individual component signals. The proposed approach does not require an explicit estimation of the noise level or a priori knowledge of the signal-to-noise ratio as is usually needed in most common speech-enhancement methods. The effectiveness of the proposed method has been assessed based on over 1000 speech recordings from the public Librispeech database. The research included various types of noise (among others white, violet, brown, blue, and pink) and various types of disturbance (among others traffic sounds, hair dryer, and fan), which were added to the selected test signals. The score of perceptual evaluation of speech quality, allowing for the assessment of the quality of enhanced speech, and signal-to-noise ratio, allowing for the assessment of the effectiveness of disturbance attenuation, are selected for the evaluation of the resultant effectiveness of the proposed approach. The resultant effectiveness of the proposed approach is compared with other selected speech-enhancement methods or denoising techniques available in the literature. The experimental research results show that the proposed method performs better than conventional methods in many types of high-noise conditions in terms of producing less residual noise and lower speech distortion.

Keywords: adaptive thresholds; enhanced empirical wavelet transform; denoising; speech enhancement; Teager energy operator

Citation: Kuwałek, P.; Jęśko, W. Speech Enhancement Based on Enhanced Empirical Wavelet Transform and Teager Energy Operator. *Electronics* **2023**, *12*, 3167. https://doi.org/10.3390/electronics12143167

Academic Editor: Chiman Kwan

Received: 22 June 2023
Revised: 18 July 2023
Accepted: 19 July 2023
Published: 21 July 2023

1. Introduction

Noise has been a problem since the first sound recording devices, such as the Edison phonograph, were developed. These devices recorded voices and other sounds. Unfortunately, while listening to the recordings after information acquisition, one could always observe various unpleasant disturbances occurring in them. Some were caused by the environment, and others (e.g., crackling) were the result of very poorly technologically advanced devices [1]. By the end of the 20th century, technological advances had minimized the mechanical factors of devices causing an obvious signal disturbance. Despite this, noise is still present in modern audio systems, albeit at a low level. All recordings, even those made in hermetic rooms, contain some background noise that is picked up by the microphone. Contemporary microphones have high sensitivity, which results in better clarity and accuracy of the recorded information [2]. Another factor that causes noise in an audio recording is that each part of the recording equipment (e.g., a microphone, amplifier, or mixer that processes the recorded signal) adds a certain electronic noise. Its occurrence in the recording is the result of electric fluctuations that occur as a result of the

chaotic movement of charge in electronic components [3]. One of the types of such noise is thermal noise [4]. In addition, disturbance can also be caused by low-level magnetic and electrostatic fields in and around buildings from supply circuits. The method to limit this is to use shielded cables with appropriately designed insulation, most often with XLR connectors [5].

Noise reduction is still under investigation in modern audio engineering. The development of semiconductor technology has contributed to the progress of miniaturization, which, in the area of computer technology, has allowed for the collection of more information and faster processing. Currently, recordings of audio signals can be made using a wide range of available devices. The choice of a specific tool is made mainly in terms of the purpose of registrations and considering the conditions under which the registration is carried out. Presently, disturbances affecting the useful signal can have various characteristics. In an open space, inexpedient sounds can be road and industrial noises, while indoors, air conditioning, household appliances, office machines, or other general-use devices can be noise sources. It happens that the sounds that make daily life easier can also be treated under certain conditions as a disturbance degrading the signal, e.g., loud music. In particular, this problem can be noticed in the currently widely developed voice recognition and monitoring systems, where additional noises can have a negative impact on the correctness of detection, verification, identification, or recognition. Hence, it is important to use various speech-enhancement (noise-reduction) techniques. In the context of speech or speaker recognition based on the pattern recognition process that separates analysis from recognition, speech enhancement already has great potential. Speech enhancement can also be used in coding and with a variety of devices such as audio prostheses.

Considering the processing domain, traditional speech-enhancement methods can be divided into three categories, namely time [6–10], frequency [11–14] and time–frequency [15–18] domain methods. Despite the significant development of speech-enhancement techniques [19–25] from particular categories in recent years, there is still a problem of obtaining a high-quality speech signal with high noise attenuation. Typically, if a method maintains high noise attenuation properties, such a method does not always provide a significant improvement in speech quality. On the other hand, many methods with the potential to significantly increase the quality of speech do not allow high attenuation of the noise signal. The abovementioned problem significantly worsens in the case of highly noisy speech signals. Therefore, considering the presented problems, it can be seen that in recent years there has been a growing tendency to solve speech-enhancement problems using machine-learning methods [26–29], in particular using various deep-learning techniques [30–34]. However, machine-learning approaches, compared to traditional speech-enhancement techniques, often require significant computational capabilities of the hardware on which the approach is implemented, and it is necessary to create a representative dataset that includes possible numerous real-world cases. It is worth noting that despite the creation of a representative training and validation dataset, the over-fitting of the appropriate model can still occur, which cannot be generalized. Therefore, methods of traditional speech enhancement are still being developed, regardless of machine-learning methods. Hence, the paper presents a new traditional method of speech enhancement based on a new high-efficiency decomposition technique for non-stationary noisy signals, i.e., an enhanced empirical wavelet transform [35], which, in combination with the Teager energy operator [36], allows the high suppression of the noise signal (based on the signal-to-noise ratio (SNR) [37]) and allows for a significant improvement in the quality of speech (based on the perceptual evaluation of speech quality (PESQ) measure [38]), especially in the case of a highly noisy signal of speech. The proposed approach does not require an explicit estimation of the noise level or a priori knowledge of the signal-to-noise ratio, as is usually needed in most common speech-enhancement methods. The presented approach was verified and compared with other available approaches in the literature. Verifications of the proposed approach in the paper were carried out on recordings from one microphone. It is worth noting that the proposed approach has the possibility of multidimensional analysis,

and thus, it is possible to apply/extend the proposed approach to the analysis of speech signals from many microphones.

2. Proposed Approach

In the proposed approach shown in Figure 1, the processed speech signal $x(t)$ is first decomposed into a maximum of N component signals $x_{\mathrm{EEWT}_k}(t)$ using the enhanced empirical wavelet transform (EEWT) [35]. Each component signal $x_{\mathrm{EEWT}_k}(t)$ is successively processed according to the following steps:

1. Determination of Teager Energy Operator $x_{\mathrm{TEO}_k}(t)$ [36];
2. Mask construction $M_k(t)$ based on $x_{\mathrm{TEO}_k}(t)$;
3. Mask processing $M'_k(t)$;
4. Determination of the time–space adaptation of thresholds $\lambda'_k(t)$ based on $M'_k(t)$;
5. Realization of soft thresholding $x'_{\mathrm{EEWT}_k}(t)$ considering $\lambda'_k(t)$ for the component signal $x_{\mathrm{EEWT}_k}(t)$.

In the last step, the individual processed signals $x'_{\mathrm{EEWT}_k}(t)$ are summed, as a result of which an enhanced speech signal $x'(t)$ is obtained.

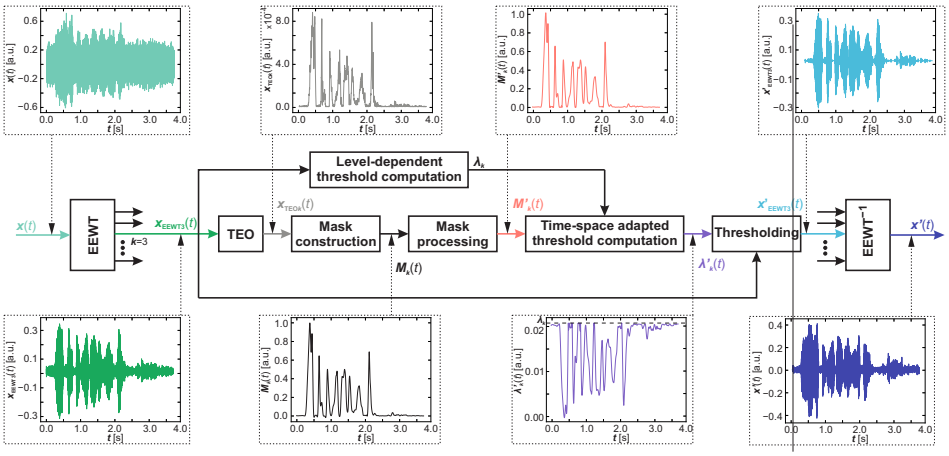


Figure 1. Diagram of the proposed speech-enhancement approach.

2.1. Enhanced Empirical Wavelet Transform (EEWT)

The EEWT algorithm can be represented in the following steps.

1. The use of fast Fourier transform (FFT) to determine the spectrum of the analyzed signal.
2. The calculation of the upper envelope of the analyzed signal using the l-th order statistical filter (OSF). In the enhanced method [35] (in relation to the conventional empirical wavelet transform (EWT) [39]), the envelope is used to identify the trend of spectrum variation. The number of order l is determined according to the relationship:

$$l = \lfloor n_{\mathrm{tap}} \cdot \frac{L}{f_s} \rfloor, \tag{1}$$

where n_{tap} is the filter order scaling parameter ($n_{\mathrm{tap}} = 70$ was adopted in the research—see Section 2.6), L is the length of the analyzed signal, and f_s is the sampling frequency of the analyzed signal.
3. The determination of spectrum frequency peaks from the designated envelope and the selection of useful ones based on the following criteria: (a) the width of a flat top cannot be shorter than the order statistics filter size; (b) the most representative flat top in the neighbor ones is picked out; (c) the useful flat tops do not appear in the downtrend of the analyzed signal spectrum.
4. The calculation of the spectrum segmentation boundaries based on the flat tops obtained in Step 3.

5. The construction of the empirical scaling function and empirical wavelet as in the EWT method [39], and the decomposition of the analyzed signal into component signals.

Steps 1–4 allow for the segmentation of the spectrum of the analyzed signal. To segment the spectrum, the segmentation boundaries must be determined. For this purpose, the spectrum is normalized to the range $[0; \pi]$ and is divided into up to N intervals (if the number of useful flat tops N_u is smaller than the predefined number of component signals N, then the analyzed signal is decomposed into N_u of component signals). The predefined number of component signals N of 32 was adopted in the research—see Section 2.6. The individual interval boundaries are designated as ω_k, where $\omega_0 = 0$ and $\omega_{\min(N,N_u)} = \pi$. Each sub-range is marked as $\bigcup_{k=1}^{\min(N,N_u)} \Lambda_k = [0; \pi]$. The boundary determination is based on flat tops described in Step 3. Each boundary is the minimum between subsequent flat tops in the analyzed signal spectrum. If the flat tops are designated as FT_k, then:

$$\omega_k = \arg\min_{\omega \in (FT_k, FT_{k+1})} \left(\widehat{f}(\omega) \right), \tag{2}$$

where $\widehat{f}(\omega)$ is the analyzed signal spectrum.

Step 5 allows for the construction of the empirical wavelets, allowing for the extraction of individual component signals as described in [39]. For specific intervals, an empirical scaling function Φ_k is constructed described by (3) and an empirical Meyer wavelet Ψ_k described by (4) [39]:

$$\Phi_k = \begin{cases} 1 & \text{if} \quad |\omega| \leq \omega_k - \tau_k \\ \cos\left[\frac{\pi}{2} v\left(\frac{1}{2\tau_k}(|\omega| - \omega_k + \tau_k)\right)\right] & \text{if} \quad \omega_k - \tau_k < |\omega| < \omega_k + \tau_k \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

$$\Psi_k = \begin{cases} 1 & \text{if} \quad \omega_k + \tau_k \leq |\omega| \leq \omega_{k+1} - \tau_{k+1} \\ \cos\left[\frac{\pi}{2} v\left(\frac{1}{2\tau_{k+1}}(|\omega| - \omega_{k+1} + \tau_{k+1})\right)\right] & \text{if} \quad \omega_{k+1} - \tau_{k+1} < |\omega| < \omega_{k+1} + \tau_{k+1} \\ \sin\left[\frac{\pi}{2} v\left(\frac{1}{2\tau_k}(|\omega| - \omega_k + \tau_k)\right)\right] & \text{if} \quad \omega_k - \tau_k \leq |\omega| \leq \omega_k + \tau_k \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $v(t)$ can be described as:

$$v(t) = \begin{cases} t^4(35 - 84t + 70t^2 - 20t^3) & \text{if} \quad 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

For the defined empirical wavelet, the τ_k can be selected in many ways and determines the appropriate width of the spectrum segment. One of the simplest choices is τ_k proportional to ω_k, so $\tau_k = \zeta\omega_k$, where $0 < \zeta < 1$.

The approximate coefficients are the scalar product of the processed signal and the empirical scaling function:

$$W_x^{\varepsilon}(0, t) = \langle x, \Phi_1 \rangle = \int x(\tau)\overline{\Phi_1(\tau - t)}d\tau. \tag{6}$$

The detail coefficients are the scalar product of the processed signal and the empirical wavelet:

$$W_x^{\varepsilon}(k, t) = \langle x, \Psi_k \rangle = \int x(\tau)\overline{\Psi_k(\tau - t)}d\tau. \tag{7}$$

For the defined approximation coefficients and detail coefficients, signal decomposition defined by empirical wavelet transform can be described as:

$$x_{\mathrm{EEWT}_k}(t) = \begin{cases} W_x^{\varepsilon}(0,t) * \Phi_1(t) & \text{if} \quad k = 0 \\ W_x^{\varepsilon}(k,t) * \Psi_k(t) & \text{if} \quad k = 1,2,\ldots,\min(N,N_u) \end{cases}. \tag{8}$$

In the research, the software provided by the authors of the paper [40] was used to prepare the enhanced EWT method. A method that allows for the calculation of the upper envelope of the analyzed signal using the specific order statistical filter was added to the available software. In addition, based on the obtained envelope, a method was added that allows for the determination of spectrum frequency peaks and selection of useful ones based on the criteria characteristic of EEWT.

2.2. Teager Energy Operator (TEO)

For the time–space-adapted threshold construction, the Teager energy operator (TEO) [36] is used, which for the k-th component signal $x_{\mathrm{EEWT}_k}(t)$ (in the discrete domain for sampling moments t_i) is given the following relationship:

$$x_{\mathrm{TEO}_k}(t_i) = x_{\mathrm{EEWT}_k}(t_i) \cdot x_{\mathrm{EEWT}_k}(t_i) - x_{\mathrm{EEWT}_k}(t_{i+1}) \cdot x_{\mathrm{EEWT}_k}(t_{i-1}). \tag{9}$$

The use of the Teager energy operator allows for the increase of the ability to recognize a speech signal from a disturbance (noise) signal.

2.3. Mask Construction

The mask construction for the k-th component signal is performed by smoothing the signal $x_{\mathrm{TEO}_k}(t)$ (obtained with the Teager energy operator (TEO)) and then is normalized. Thus, the initial construction of the mask $M_k(t)$ is given by the relationship:

$$M_k(t) = \frac{x_{\mathrm{TEO}_k}(t) * h_k(t)}{\max\left(\left|x_{\mathrm{TEO}_k}(t) * h_k(t)\right|\right)}, \tag{10}$$

where $h_k(t)$ is the impulse response for a second-order Butterworth low-pass filter (infinite impulse response (IIR) filter).

2.4. Mask Processing

The time-scale adaptation of thresholds should be adjusted to the speech waveform; therefore, the difference between the local maxima should be reduced. The processing result should provide a compromise between noise attenuation and speech distortion. Thus, the pre-calculated mask is processed according to the relationship:

$$M'_k(t) = \begin{cases} \left(\frac{|M_k(t)| - S_k}{\max(|M_k(t)| - S_k)}\right)^{\gamma} & \text{if} \quad S_k < S_{\mathrm{thres}} \\ 0 & \text{if} \quad S_k \geq S_{\mathrm{thres}} \end{cases}, \tag{11}$$

where S_k is the parameter called offset and is determined as the abscissa of the maximum of the amplitude distribution H of the corresponding mask $M_k(t)$, and is estimated over the analyzed frame:

$$S_k = \mathrm{abscissa}[H(M_k(t))], \tag{12}$$

γ is a parameter that affects the compromise between noise attenuation and speech distortion ($\gamma = 0.125$ was adopted in the research—see Section 2.6); S_{thres} is a parameter that affects the ability to discern speech from silence ($S_{\mathrm{thres}} = 0.041$ was adopted in the research—see Section 2.6).

2.5. Thresholding

As part of the thresholding, the level-dependent threshold is first determined in accordance with the dependence:

$$\lambda_k = \sigma_k \frac{2 \log L}{\log_2(k+2)}, \tag{13}$$

where $k = 0, 1, 2, \ldots, \min(N, N_u)$, $\sigma_k = MAD_k / 0.6745$ is the noise level, L is the length of the analyzed signal, MAD_k is the median absolute deviation value for $x_{\mathrm{EEWT}_k}(t)$.

In the next step, based on the level-dependent threshold λ_k and the processed mask $M'_k(t)$, the time–space adaptation of thresholds $\lambda'_k(t)$ is determined according to the dependence:

$$\lambda'_k(t) = \begin{cases} \Gamma\lambda_k(1 - M'_k(t)) & \text{if} \quad \frac{P_k}{P} < P_{\mathrm{thres}} \\ \lambda_k(1 - M'_k(t)) & \text{if} \quad \frac{P_k}{P} \geq P_{\mathrm{thres}} \end{cases}, \tag{14}$$

where P_k is the power of the k-th component signal $x_{\mathrm{EEWT}_k}(t)$, P is the power of the processed speech signal $x(t)$, P_{thres} is the level-dependent threshold λ_k scaling trigger ($P_{\mathrm{thres}} = 0.170$ was adopted in the research—see Section 2.6); Γ is the scaling parameter of the level-dependent threshold λ_k ($\Gamma = 4.837$ was adopted in the research—see Section 2.6).

In the last step, soft thresholding [41] is performed for individual k-th component signals $x_{\mathrm{EEWT}_k}(t)$ based on the determined time–space adaptation of thresholds $\lambda'_k(t)$.

2.6. Final Processing

To obtain an enhanced speech signal $x'(t)$, the individual processed k-th component signals $x'_{\mathrm{EEWT}_k}(t)$ are summed according to the relationship:

$$x'(t) = \sum_{k=0}^{\min(N, N_u)} x'_{\mathrm{EEWT}_k}(t). \tag{15}$$

It is worth noting that the set of hyperparameters adopted in the research for proposed approach—$[n_{\mathrm{tap}}, N, S_{\mathrm{thres}}, \gamma, P_{\mathrm{thres}}, \Gamma]$—is the result of the optimization process carried out using Monte Carlo analysis for a subset of the selected test signal database with added "white" noise with different SNR_u values. The optimization process was focused on maximizing speech quality. The optimization process included many combinations of hyperparameters randomly selected from the space bounded by a hyperplane: $[10, 3, 0, 0.125, 0.01, 1]$ and $[500, 100, 1, 3, 0.3, 5]$. The effectiveness of the proposed approach will likely increase if different types of disturbances are included in the optimization process. Nevertheless, by comparing the proposed approach with the adopted hyperparameter values with other speech-enhancement methods available in the literature, better effectiveness is obtained.

3. Materials

In the carried-out research, the well-known, publicly available Librispeech [42] database was used as a source of clean speech. The Librispeech database is a large-scale corpus of read speech in English with 1000 h of recordings. Its structure is divided into separate subsets, which, by default, are used to train, develop, and evaluate automatic speech recognition systems ("train", "dev", and "test" sets, respectively). Due to the extensive collection of recordings in the Librispeech database, only the test subset was used in the experiments, or more precisely the "test-clean" set. This collection contains a total of approx. 2600 speech files, the duration of which ranges from 1 to 35 s (a total of over 5 h of recordings), while approx. 400 recordings are equal to or shorter than 3 s, 800 recordings are equal to or shorter than 4 s, 1000 recordings are equal to or shorter than 5 s, and 2000 recordings are equal to or shorter than 10 s. Initial experiments were shown that consider that each of the methods of speech enhancement or noise reduction described in this paper is time-consuming, even on a limited, selected subset of the Librispeech database ("test-clean"). Therefore, it was

decided to use a smaller but representative number of recordings and speakers from the subset of the "test-clean" set, in which recordings with a duration of no longer than 5 s (exactly 1093 speech files; 1 h of recordings) in total, including each of (i.e., 40) speakers from the "test-clean" set), were used. Speech files selected in this way from the Librispeech database were used as a reference set of clean speech, to which were added various types of selected noises and ambient sounds. Speech files prepared in such a way, containing noise, were used as the input data for each of the tested methods of speech enhancement or noise reduction.

4. Results and Discussion

The research used the database of speech signals described in Section 3. For the assumed database of test speech signals, the following types of noise [43] were added successively: white, pink, blue, violet, and brown. The following types of disturbances: traffic noise, fan, a hair dryer (ambient sounds recordings from own resources) with SNR_u values equal to 20 dB, 10 dB, 0 dB, −10 dB, −20 dB, where SNR_u is unprocessed signal-to-noise ratio determined according to the relationship:

$$\mathrm{SNR}_u = 10 \log \frac{\sum_{i=1}^{L} \left\{ \left(x_{\mathrm{clean}}(t_i) \right)^2 \right\}}{\sum_{i=1}^{L} \left\{ \left(x_{\mathrm{noise}}(t_i) \right)^2 \right\}}, \tag{16}$$

where $x_{\mathrm{clean}}(t_i)$ is the non-distorted (clean) speech signal, $x_{\mathrm{noise}}(t_i)$ is the added noise (disturbance) signal, t_i is the sampling moments, L is the length of the analyzed signal. Color noise is generated by taking uniform white noise and filtering with a coloring filter to obtain the desired noise spectrum with a power spectral density function given by:

$$S(f) = \frac{\mathcal{L}(f)}{|f|^{\alpha}}, \tag{17}$$

where α is a real number in the interval $[-2, 2]$ and $\mathcal{L}(f)$ is a positive, slowly varying or constant function. Pink noise ($\alpha = 1$) has equal energy per octave and the power spectral density of pink noise decreases 3 dB per octave. Blue noise ($\alpha = -1$) has a power spectral density that increases by 3 dB per octave. Violet noise ($\alpha = -2$) has a power spectral density that increases by 6 dB per octave. Brown noise ($\alpha = 2$) has a power spectral density that decreases by 6 dB per octave.

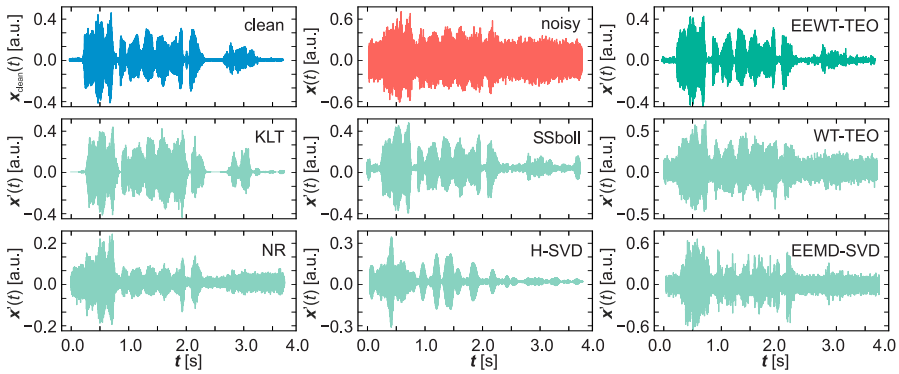
For individual considered noisy speech signals, the process of speech enhancement was implemented using the proposed approach and selected methods of noise reduction or speech enhancement available in the literature, i.e., Karhunen–Loeve transform (KLT)—a generalized subspace approach of KLT [44,45], SSboll—spectral subtraction [13,46], WT-TEO—wavelet speech enhancement based on time-scale adaptation [47], NR—noise reduction using spectral gating [48,49], H-SVD—noise-reduction technique based on Hankel matrix and singular value decomposition [50], EEMD-SVD—noise-reduction technique based on ensemble empirical mode decomposition and singular value decomposition [51].

For the individual speech-enhancement results obtained by the considered methods for the considered noisy speech signals, two measures were determined, i.e., the signal-to-noise ratio (SNR) [37] assessing the ability to suppress noise (disturbance) by the considered methods, and the perceptual evaluation of speech quality (PESQ) score [38] assessing the quality of the processed speech signal. The PESQ score was determined using the available Python library [52], and the SNR was calculated according to the relationship [37]:

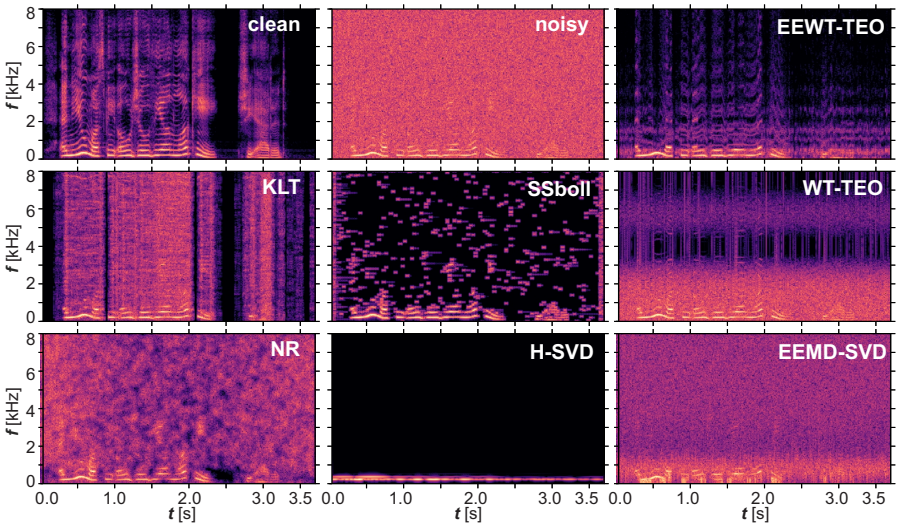
$$\mathrm{SNR} = 10 \log \frac{\sum_{i=1}^{L} \left\{ \left(x_{\mathrm{clean}}(t_i) \right)^2 \right\}}{\sum_{i=1}^{L} \left\{ \left(x'(t_i) - x_{\mathrm{clean}}(t_i) \right)^2 \right\}}, \tag{18}$$

where $x_{\text{clean}}(t_i)$ is the clean signal without added noise, and $x'(t)$ is the result of the considered methods of speech enhancement or noise reduction.

Figure 2a shows a comparison of the effectiveness of the considered methods for the selected speech signal with the "white" noise with the value of $\text{SNR}_u = 0\,\text{dB}$ in the time domain. Based on the presented selected signals in the time domain, it can be seen that the best representation of the original signal (e.g., by comparing the envelope signals) is obtained for the proposed approach as well as for KLT and SSbol.



(**a**) time domain



(**b**) time–frequency domain

Figure 2. The comparison of the effectiveness of the considered methods for the selected speech signal with the "white" noise with the value of $\text{SNR}_u = 0\,\text{dB}$ in the time (**a**) and the time–frequency (**b**) domain.

Figure 2b shows a comparison of the effectiveness of the considered methods for the selected speech signal with the "white" noise with the value of $\text{SNR}_u = 0\,\text{dB}$ in the time–frequency domain with the use of spectrograms. Based on the presented selected spectrograms, it can be seen that the best recreation of signal features in the time–frequency domain is obtained for the proposed approach. This property is especially important for the speech recognition process [53–57] in which deep learning is used. In such a process, algorithms can learn from features that are not always relevant to the listener. Thus, obtaining a high PESQ/SNR and recreation of the appropriate envelope cannot always be sufficient. Hence, a new approach has been proposed, which also allows for the reconstruction of important features in the time–frequency domain, thanks to which the use of such a block at the beginning of the signal chain in the speech recognition process allows for a significant increase in the accuracy of speech recognition in a situation when

the speech signal is noisy or disturbed, which is the subject of a separate publication. The use of other selected speech-enhancement methods in the process of speech recognition based on deep learning, in most cases deteriorated recognition, is due to the removal of features that do not have to be important from the point of view of the listener, but are important from the point of view of the deep-learning system.

Figures 3–7 show a comparison of the distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals, considering "white", "violet", "brown", "blue", and "pink" noise, respectively. The distribution of the obtained results is presented using the boxplot. Due to the readability of the presented research results, the mean value for individual boxplots is not plotted, because it almost always fluctuated around the median or coincided with the median.

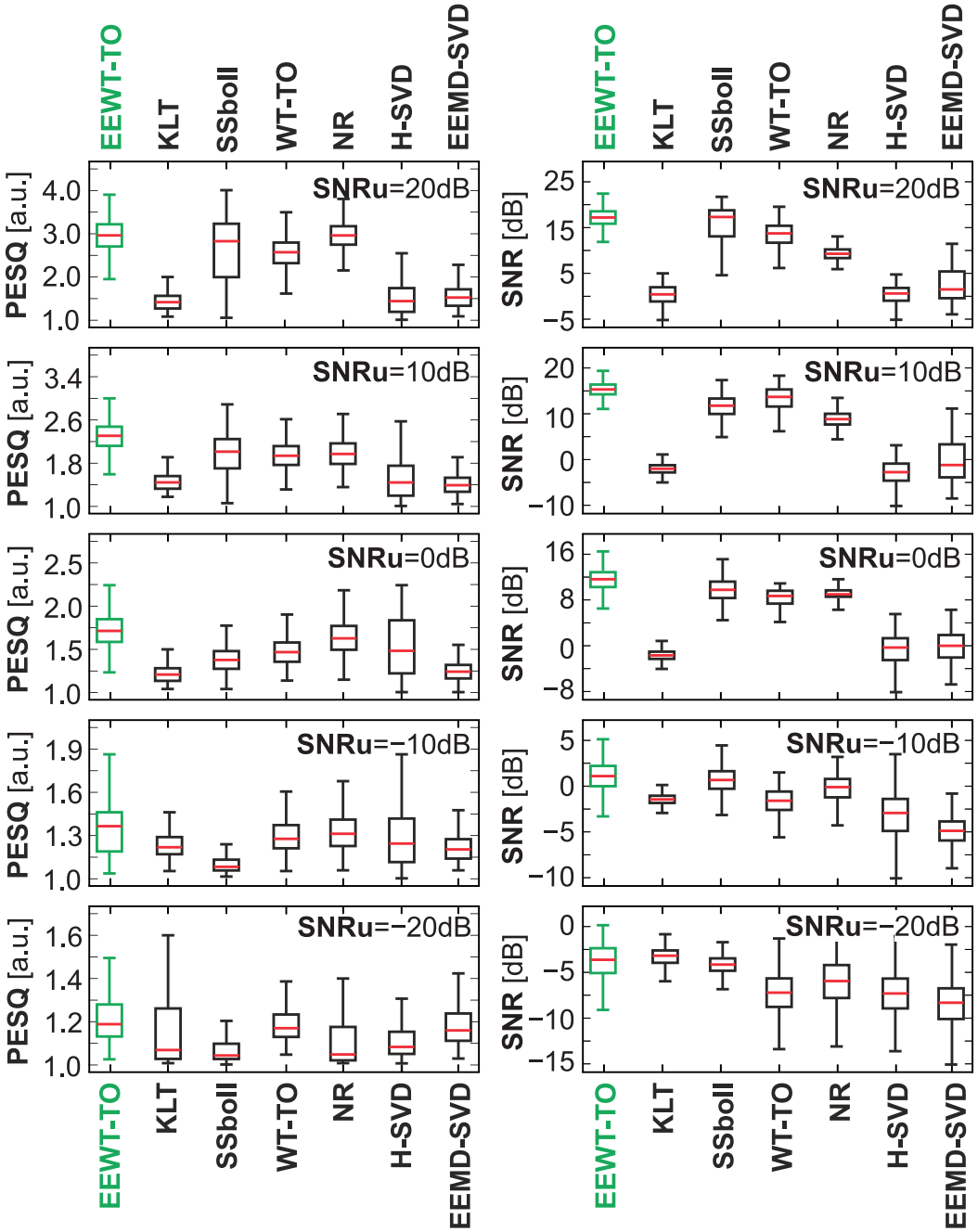


Figure 3. The distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals with "white" noise.

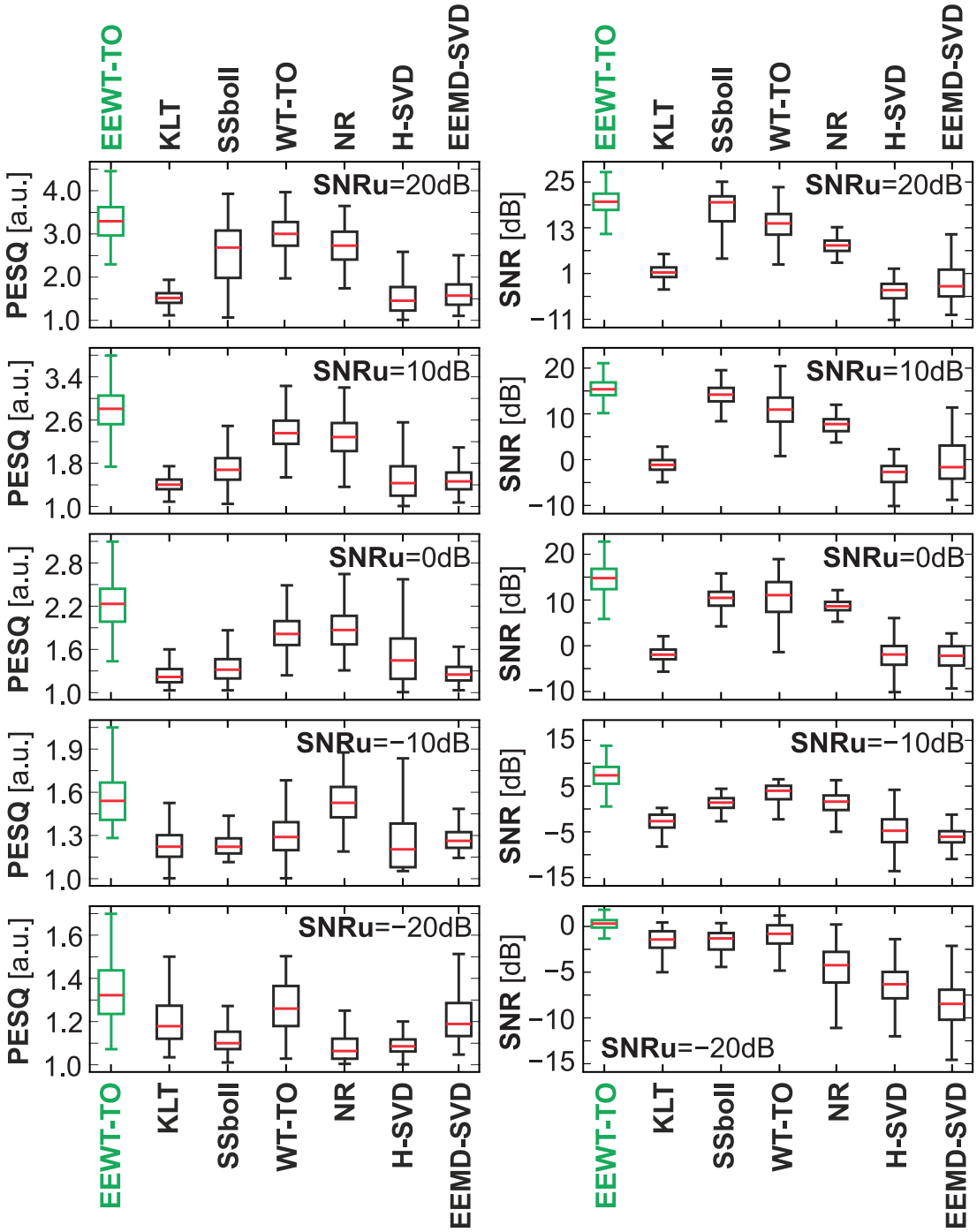


Figure 4. The distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals with "violet" noise.

Figures 8–10 show a comparison of the distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals, considering disturbances such as traffic sounds, hair dryer, and fan. The distribution of the obtained results is presented using the boxplot. Due to the readability of the presented research results, the mean value for individual boxplots is not plotted, because it almost always fluctuated around the median or coincided with the median. Table 1 shows the maximum and median values of PESQ score for individual considered cases. Table 2 shows the maximum and median values of the SNR for individual considered cases.

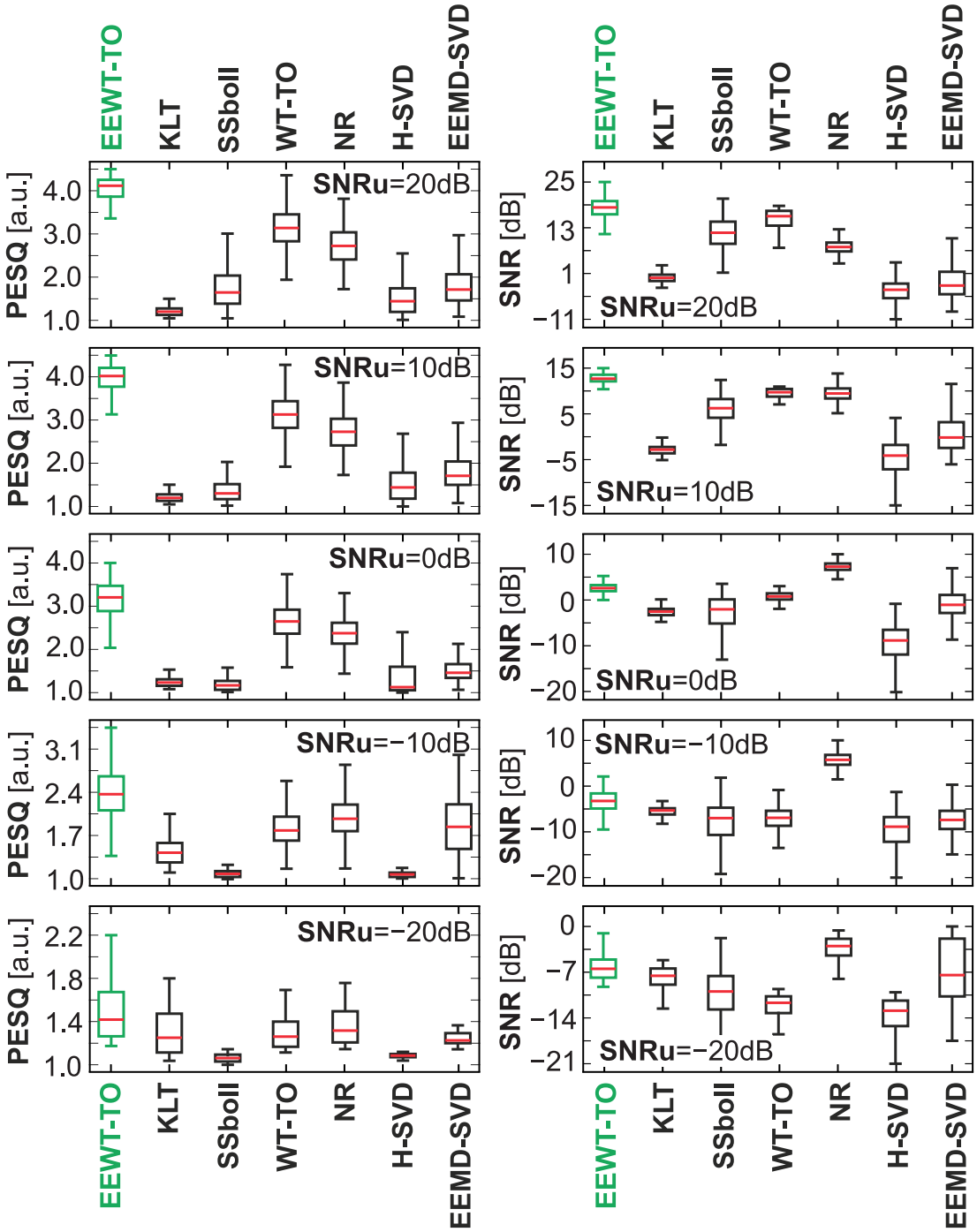


Figure 5. The distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals with "brown" noise.

Based on the research results presented in Figures 3–10 regarding the PESQ score, it can be seen that the proposed approach provides the best global enhancement of speech quality for various types of noise and various types of disturbances (ambient sounds). The greatest effectiveness of the proposed approach is visible for the SNR_u at the level from 20 dB to 0 dB.

Based on the research results presented in Figures 3–10 regarding the SNR, it can be seen that the proposed approach provides the best noise (disturbance) attenuation for the analyzed speech signals for various types of noise and various types of disturbances (ambient sounds). In the case of this parameter, the high efficiency of the proposed approach is evident for both a low and a high SNR_u value. Additionally, it is worth noting that

the dispersion of the research results for the proposed approach is small in the case of SNR, i.e., the resultant mean value, median value, lower and upper quartile, and the minimum and maximum value excluding outliers, and almost coincides with the adopted axis scale considering the dispersion for each analyzed speech enhancement or noise-reduction methods. On this basis, it can be concluded that the high noise (disturbance) attenuation ability will be maintained with a high probability for various types of other noises (disturbances). In the case of other considered methods, the dispersion is much more noticeable, which indicates significantly changeable effectiveness.

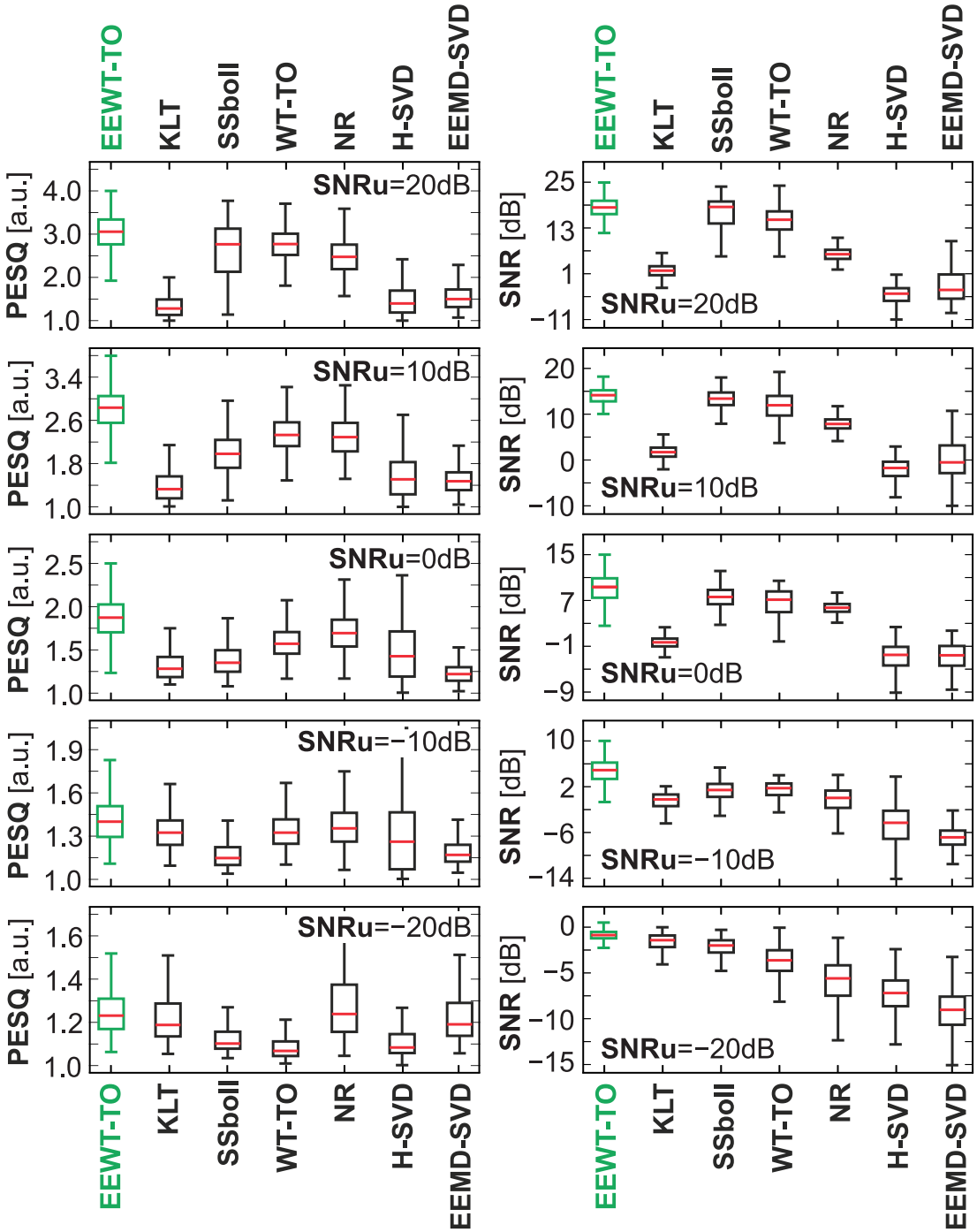


Figure 6. The distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals with "blue" noise.

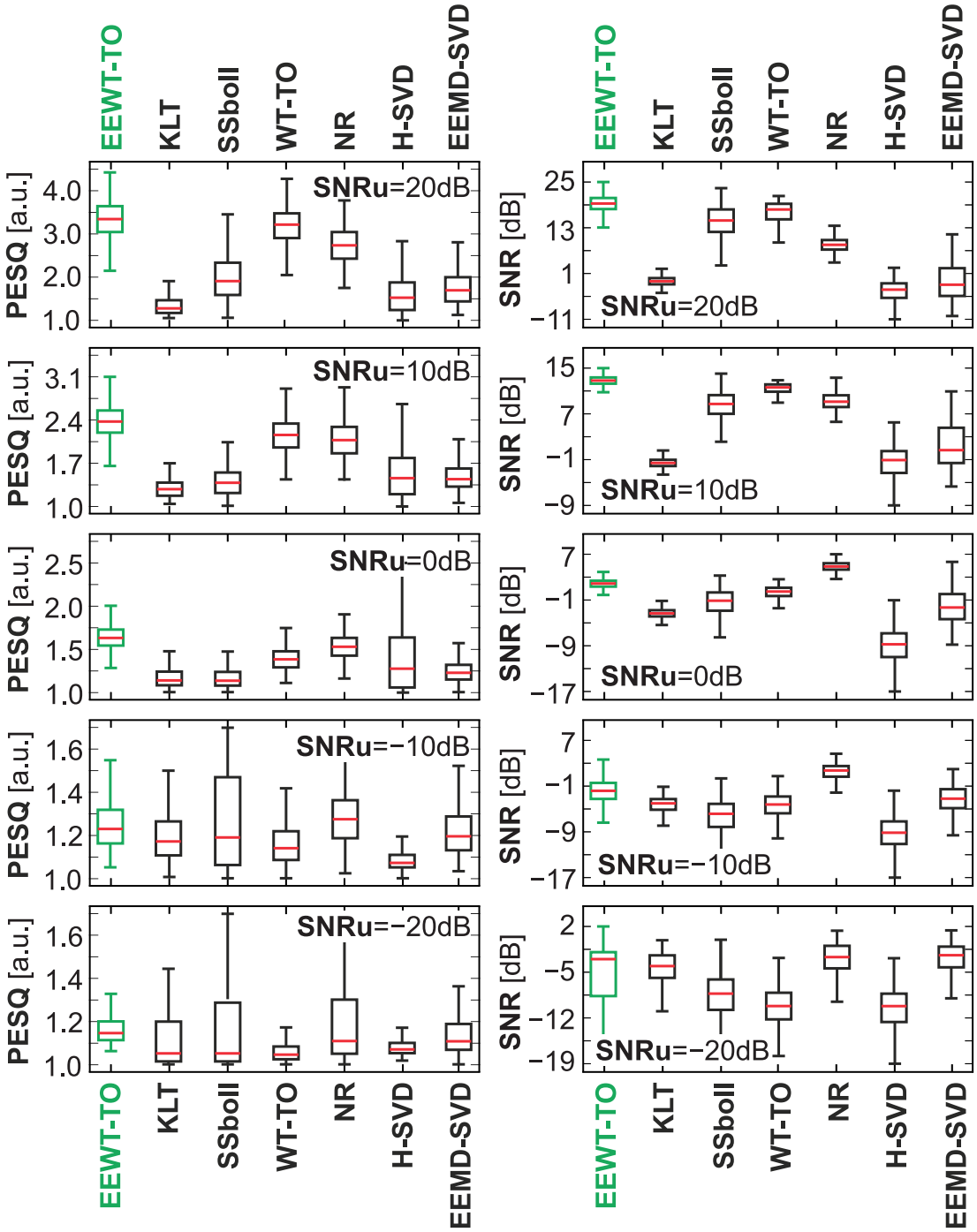


Figure 7. The distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals with "pink" noise.

The highest efficiency is understood as maintaining the highest median value and the mean value of PESQ and SNR for various types of disturbing signals with different SNR_u values. In the case of SNR, both the aforementioned median and mean values as well as the other statistical parameters (e.g., minimum and maximum value) are the highest among all the methods considered in the research. In the case of PESQ, other methods have sometimes obtained higher maximum PESQ values (e.g., SSboll for "pink" noise with $SNR_u = -20$ dB) or higher minimum PESQ values (e.g., KLT for "white" noise with $SNR_u = -10$ dB). This situation means that sometimes, as a result of the noise (disturbance) attenuation, the speech quality can be distorted from the listener's point of view, or sometimes it is observable to obtain a speech signal of better quality for other methods than the proposed

approach. It is worth noting that extending the process of optimization of hyperparameters (see Section 2.6), considering other selected noises, could make better speech-enhancement results for the proposed method in selected cases for noise other than "white" noise, where other methods provided better results (e.g., "pink").

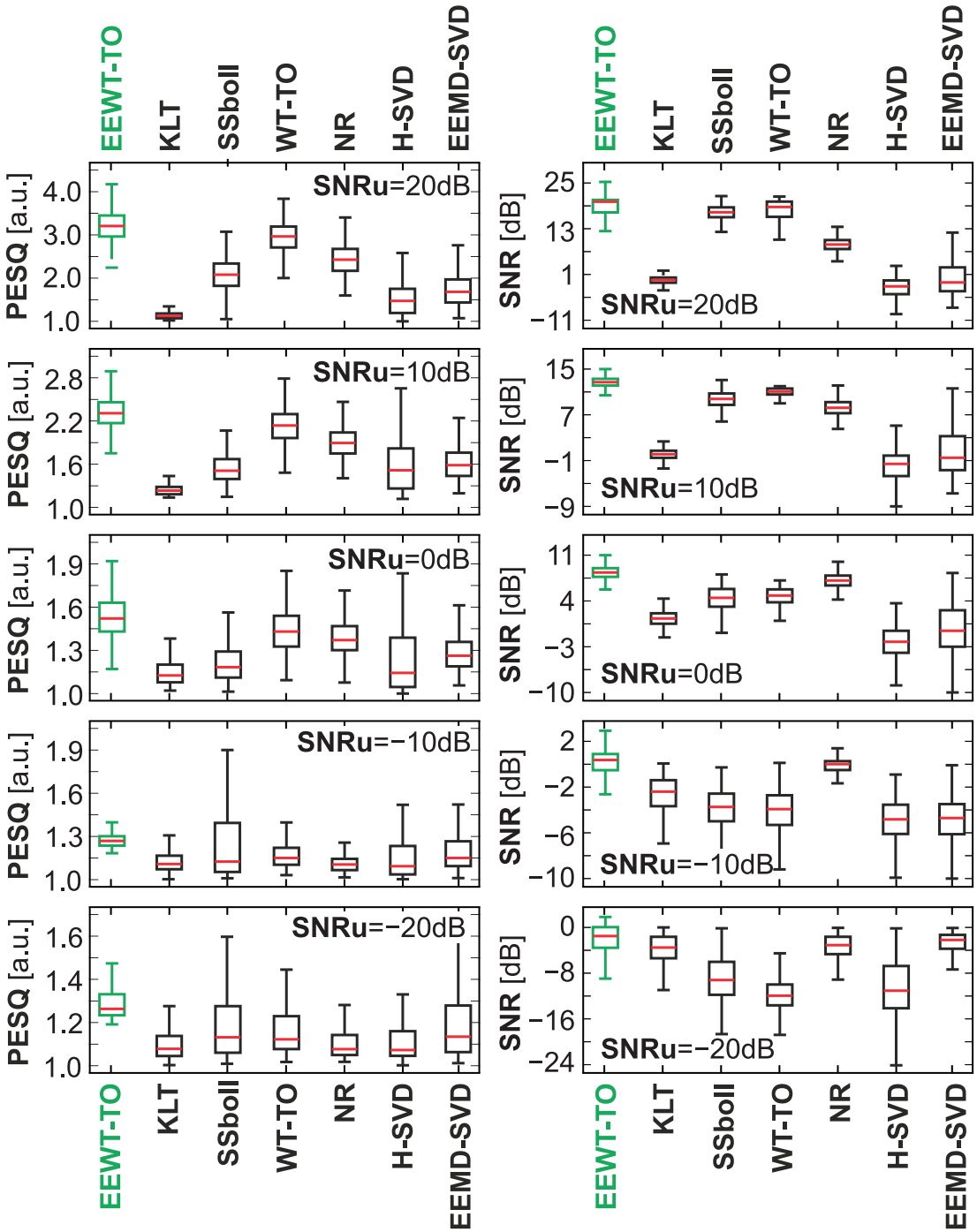


Figure 8. The distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals with ambient sound like "traffic sounds".

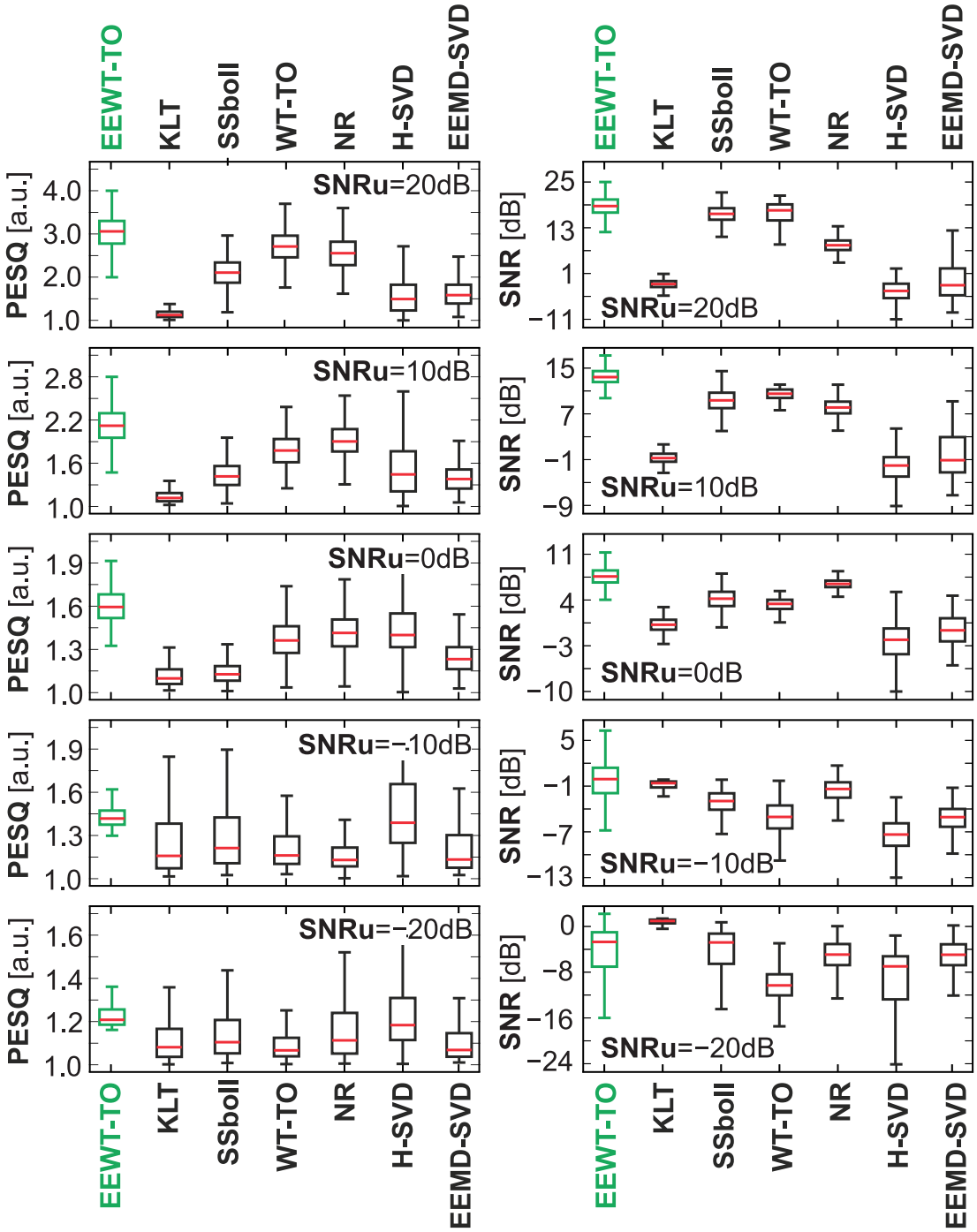


Figure 9. The distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals with ambient sound like "hair dryer".

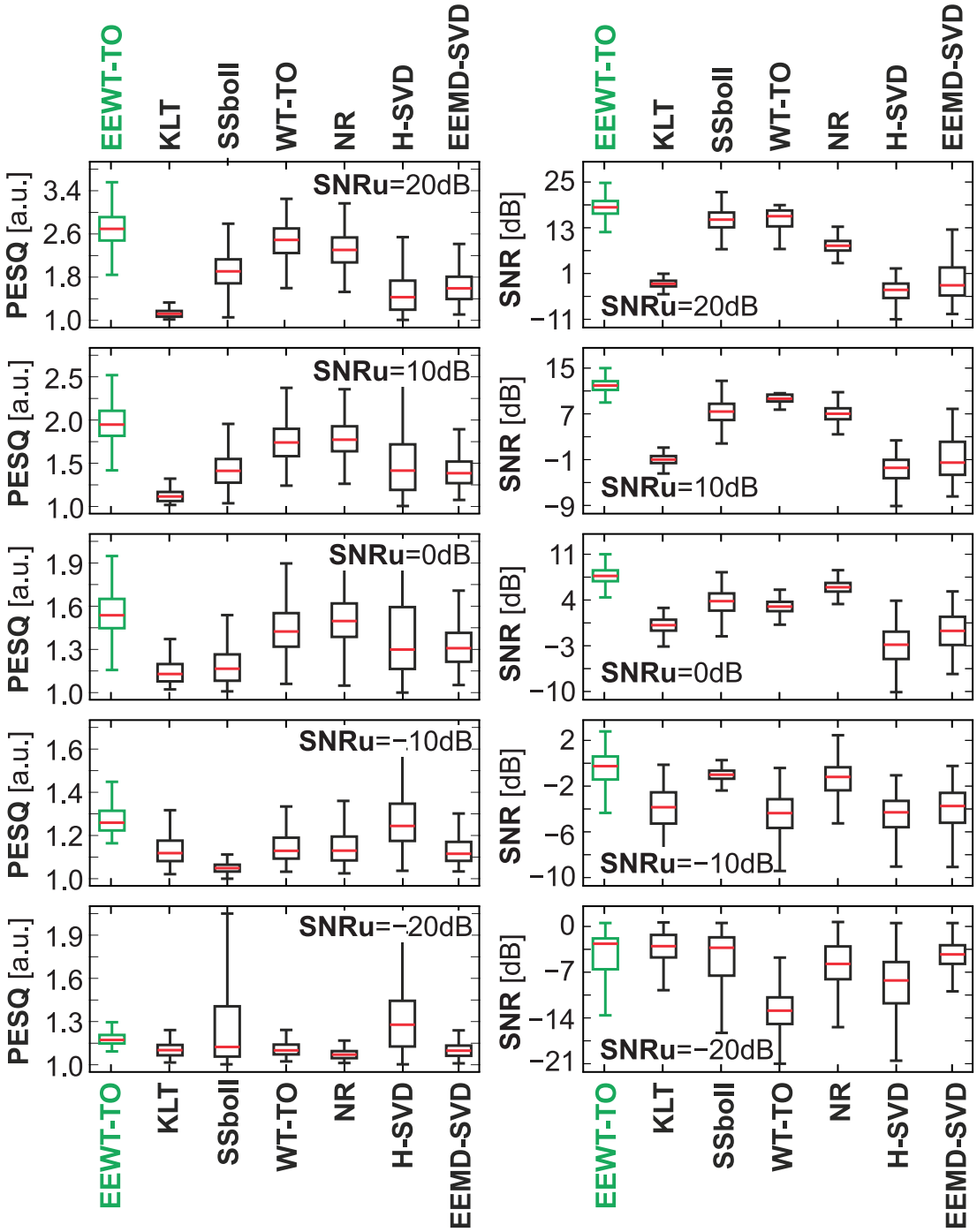


Figure 10. The distribution of PESQ and SNR results for the proposed approach and other considered methods for analyzed speech signals with ambient sound like "fan".

Table 1. The maximum and median values of PESQ score for individual considered cases.

Disturbance Type	SNRu	The Maximum Value							The Median Value						
		EEWT-TO	KLT	SSboll	WT-TO	NR	H-SVD	EEMD-SVD	EEWT-TO	KLT	SSboll	WT-TO	NR	H-SVD	EEMD-SVD
"white" noise	20	3.90	2.00	4.01	3.50	3.81	2.55	2.28	2.97	1.42	2.82	2.57	2.96	1.44	1.51
	10	3.00	1.91	2.89	2.61	2.71	2.57	1.91	2.31	1.44	2.01	1.93	1.96	1.47	1.40
	0	2.24	1.50	1.77	1.90	2.18	2.24	1.55	1.71	1.21	1.37	1.45	1.63	1.48	1.24
	−10	1.86	1.46	1.24	1.61	1.68	1.87	1.48	1.36	1.22	1.10	1.27	1.31	1.27	1.19
	−20	1.49	1.60	1.20	1.39	1.40	1.31	1.42	1.19	1.07	1.05	1.18	1.05	1.08	1.16
"violet" noise	20	4.45	1.94	3.93	3.97	3.65	2.58	2.51	3.29	1.51	2.68	3.03	2.74	1.46	1.57
	10	3.78	1.75	2.49	3.23	3.20	2.56	2.09	2.79	1.40	1.69	2.36	2.28	1.43	1.46
	0	3.10	1.60	1.86	2.49	2.65	2.57	1.64	2.23	1.24	1.30	1.81	1.87	1.47	1.26
	−10	2.05	1.53	1.44	1.68	1.88	1.84	1.48	1.54	1.22	1.15	1.29	1.53	1.23	1.26
	−20	1.70	1.50	1.27	1.50	1.25	1.20	1.51	1.32	1.18	1.10	1.26	1.08	1.09	1.19
"brown" noise	20	4.50	1.50	3.01	4.36	3.81	2.55	2.97	4.11	1.20	1.71	3.14	2.73	1.44	1.71
	10	4.50	1.50	2.03	4.28	3.87	2.68	2.94	4.02	1.20	1.34	3.13	2.73	1.44	1.70
	0	4.00	1.53	1.51	3.74	3.30	2.34	2.13	3.20	1.23	1.15	2.65	2.37	1.12	1.50
	−10	3.45	2.05	1.20	2.58	2.85	1.17	3.00	2.37	1.42	1.09	1.78	1.97	1.06	1.84
	−20	2.20	1.80	1.14	1.69	1.76	1.11	1.37	1.42	1.25	1.03	1.26	1.32	1.08	1.23
"blue" noise	20	4.00	2.00	3.77	3.70	3.59	2.42	2.29	3.06	1.31	2.77	2.77	2.48	1.40	1.50
	10	3.78	2.13	2.96	3.22	3.25	2.70	2.13	2.83	1.36	1.98	2.33	2.28	1.49	1.47
	0	2.50	1.75	1.87	2.07	2.31	2.36	1.53	1.87	1.28	1.35	1.58	1.69	1.43	1.21
	−10	1.83	1.66	1.41	1.67	1.75	2.05	1.41	1.40	1.32	1.16	1.32	1.35	1.26	1.17
	−20	1.52	1.51	1.27	1.20	1.66	1.27	1.51	1.23	1.21	1.10	1.08	1.24	1.08	1.19
"pink" noise	20	4.42	1.91	3.45	4.28	3.78	2.83	2.81	3.35	1.28	1.96	3.22	2.72	1.53	1.69
	10	3.10	1.70	2.03	2.91	2.93	2.66	2.09	2.37	1.28	1.37	2.16	2.06	1.46	1.44
	0	2.00	1.47	1.47	1.75	1.91	2.49	1.57	1.63	1.14	1.16	1.39	1.51	1.28	1.23
	−10	1.55	1.49	1.70	1.42	1.62	1.19	1.52	1.24	1.17	1.19	1.14	1.27	1.08	1.20
	−20	1.33	1.43	1.69	1.17	1.65	1.17	1.36	1.15	1.05	1.05	1.06	1.11	1.08	1.11
traffic sounds	20	4.17	1.35	3.07	3.84	3.40	2.58	2.76	3.22	1.18	2.08	2.96	2.43	1.47	1.70
	10	2.89	1.44	2.07	2.79	2.47	2.65	2.24	2.31	1.23	1.51	2.14	1.88	1.52	1.58
	0	1.92	1.38	1.56	1.85	1.72	1.83	1.61	1.52	1.14	1.18	1.43	1.37	1.14	1.27
	−10	1.40	1.30	1.90	1.40	1.26	1.52	1.51	1.27	1.11	1.13	1.15	1.10	1.09	1.15
	−20	1.47	1.27	1.60	1.44	1.28	1.33	1.60	1.26	1.08	1.13	1.12	1.10	1.07	1.13
hair dryer	20	4.00	1.36	2.96	3.70	3.60	2.71	2.64	3.05	1.14	2.12	2.71	2.56	1.49	1.58
	10	2.80	1.33	1.96	2.38	2.54	2.60	1.91	2.12	1.13	1.41	1.78	1.90	1.45	1.37
	0	1.91	1.31	1.34	1.74	1.79	1.90	1.54	1.59	1.11	1.13	1.36	1.41	1.40	1.23
	−10	1.62	1.85	1.90	1.58	1.41	1.90	1.63	1.42	1.16	1.21	1.16	1.15	1.39	1.13
	−20	1.36	1.35	1.44	1.25	1.51	1.60	1.31	1.22	1.08	1.10	1.07	1.11	1.18	1.09
fan	20	3.56	1.33	2.79	3.25	3.17	2.54	2.41	2.69	1.11	1.91	2.49	2.30	1.43	1.59
	10	2.52	1.32	1.96	2.37	2.36	2.50	1.89	1.95	1.12	1.41	1.74	1.77	1.41	1.39
	0	1.95	1.37	1.54	1.90	1.97	2.05	1.71	1.55	1.13	1.17	1.42	1.49	1.30	1.30
	−10	1.45	1.32	1.11	1.33	1.36	1.60	1.30	1.26	1.12	1.04	1.13	1.13	1.26	1.13
	−20	1.30	1.24	2.05	1.24	1.17	1.92	1.24	1.18	1.09	1.12	1.10	1.06	1.28	1.09

Table 2. The maximum and median values of the SNR for individual considered cases.

Disturbance Type	SNRu	The Maximum Value							The Median Value						
		EEWT-TO	KLT	SSboll	WT-TO	NR	H-SVD	EEMD-SVD	EEWT-TO	KLT	SSboll	WT-TO	NR	H-SVD	EEMD-SVD
"white" noise	20	22.43	5.00	21.72	19.56	13.07	4.75	11.47	17.18	0.25	17.32	13.73	9.23	0.62	1.51
	10	19.38	1.12	17.36	18.32	13.47	3.13	11.14	15.29	−2.05	11.78	13.69	8.84	−2.51	−1.20
	0	16.49	0.83	15.12	10.89	11.63	5.54	6.29	11.63	−1.68	9.78	8.69	8.99	−0.60	0.02
	−10	5.14	0.13	4.46	1.50	3.20	3.50	−0.81	1.11	−1.58	0.78	−1.60	−0.25	−2.94	−3.86
	−20	0.14	−0.84	−1.71	−1.30	−1.04	−1.04	−1.97	−3.63	−3.19	−4.05	−7.16	−5.96	−7.20	−8.32
"violet" noise	20	27.63	6.06	25.07	23.65	13.10	2.21	11.25	19.82	1.28	19.45	13.80	8.25	−3.44	−2.42
	10	21.08	2.84	19.56	20.44	12.01	2.30	11.41	15.48	−1.20	14.17	11.04	7.74	−2.90	−1.64
	0	22.76	2.08	15.81	18.94	12.16	6.04	2.70	14.71	−1.95	10.35	11.06	8.63	−1.81	−2.18
	−10	13.82	0.23	4.42	6.51	6.53	4.19	−1.24	7.36	−3.53	1.39	4.00	2.01	−4.75	−6.06
	−20	1.91	0.43	0.36	1.16	0.21	−1.37	−2.08	0.14	−1.33	−1.29	−0.97	−4.15	−6.16	−8.34
"brown" noise	20	25.02	3.11	20.63	18.73	12.57	3.86	10.20	18.31	0.05	11.59	16.01	7.85	−3.35	−2.23
	10	14.98	−0.02	12.41	10.44	13.82	4.11	11.72	12.64	−3.61	6.24	9.92	9.44	−4.11	−0.04
	0	5.25	0.00	3.40	3.03	10.00	−0.83	6.95	2.59	−2.34	−2.16	1.35	7.24	−8.78	−1.08
	−10	2.10	−3.28	1.83	−0.27	10.00	−1.29	0.29	−3.26	−5.54	−7.01	−6.93	5.69	−8.88	−7.47
	−20	−1.04	−5.16	−1.78	−9.58	−0.61	−10.08	0.00	−6.49	−7.55	−9.95	−11.68	−3.00	−12.89	−7.00
"blue" noise	20	24.93	6.39	23.84	24.09	10.39	0.70	9.50	18.33	1.37	18.49	14.94	6.06	−4.30	−3.88
	10	18.20	5.58	18.03	19.23	11.76	2.93	10.73	14.13	2.70	13.42	11.97	7.88	−1.72	−0.19
	0	15.02	1.99	12.14	10.40	8.34	2.30	1.65	9.32	−0.65	7.58	7.10	5.68	−1.99	−1.58
	−10	9.98	2.08	5.35	4.01	4.06	3.77	−2.92	4.91	−0.15	1.20	1.02	0.02	−4.65	−6.85
	−20	0.48	0.01	−0.30	0.02	−1.16	−2.40	−3.25	−0.81	−1.54	−2.00	−3.62	−5.60	−7.19	−9.03
"pink" noise	20	25.02	2.18	23.40	21.33	13.50	2.44	11.25	19.21	−1.12	14.89	17.77	8.41	−3.30	−2.14
	10	15.02	0.82	14.04	12.92	13.34	5.65	10.99	12.84	−1.32	8.83	11.69	9.22	−0.84	0.90
	0	4.01	−0.90	3.38	2.84	7.00	−0.79	5.71	2.05	−3.02	−0.82	0.76	4.92	−8.15	−1.87
	−10	3.75	−0.87	0.55	0.94	4.73	−1.53	2.12	−1.54	−3.68	−5.44	−3.86	1.93	−8.59	−2.90
	−20	2.07	0.00	0.03	−2.78	1.42	−2.84	1.47	−2.98	−4.05	−8.31	−10.22	−2.58	−10.24	−2.39
traffic sounds	20	25.27	2.44	21.62	21.49	13.69	5.29	12.86	20.62	0.02	17.66	18.80	9.06	−1.94	−0.68
	10	15.01	2.15	13.07	11.96	12.08	4.96	11.56	12.74	−0.24	9.71	11.17	8.13	−1.84	−0.82
	0	11.10	4.07	7.99	7.01	10.04	3.33	8.23	8.26	0.81	4.20	4.57	7.01	−2.90	−1.13
	−10	2.93	0.00	−0.28	0.12	1.39	−0.90	0.09	0.36	−2.40	−3.74	−3.93	0.00	−4.76	−4.71
	−20	1.77	0.23	−0.23	−4.52	−0.23	−0.23	−0.23	−1.56	−3.51	−8.88	−11.78	−3.10	−10.42	−2.19
hair dryer	20	25.23	1.07	22.49	21.65	13.59	2.46	12.48	18.91	−1.65	16.87	17.80	8.60	−4.34	−1.91
	10	17.24	1.65	14.47	12.08	12.07	4.27	9.10	13.42	−0.91	9.27	10.62	7.99	−2.31	−1.35
	0	11.10	2.27	7.69	4.86	8.07	4.71	4.11	7.23	−0.97	3.63	2.79	6.01	−3.01	−1.50
	−10	6.41	−0.20	−0.20	−0.32	1.66	−2.55	−1.29	−0.20	−0.39	−3.04	−5.02	−2.16	−7.35	−5.16
	−20	2.21	1.37	0.73	−1.56	0.09	−1.59	0.17	−2.70	1.19	−2.80	−10.32	−4.99	−6.56	−4.93
fan	20	24.77	0.86	22.35	18.94	13.24	2.25	12.50	18.38	−1.77	15.10	16.01	8.16	−3.37	−2.16
	10	15.00	1.04	12.76	10.59	10.78	2.30	7.84	11.89	−1.09	7.36	10.09	6.96	−2.53	−1.60
	0	10.96	2.27	8.04	5.23	8.39	3.44	4.94	7.46	−0.56	3.37	2.49	5.60	−3.79	−1.50
	−10	2.78	−0.14	0.27	−0.42	2.45	−1.44	−0.24	−0.25	−3.85	−0.98	−4.40	−1.20	−4.29	−3.90
	−20	0.53	0.62	0.53	−4.77	0.68	0.53	0.53	−1.43	−2.89	−3.26	−12.89	−5.57	−8.27	−4.27

5. Conclusions

The paper presents speech enhancement based on the enhanced empirical wavelet transform and the Teager energy operator. The new proposed approach is less sensitive to large disturbances compared to traditional algorithms and can significantly reduce noise as well as other disturbing signals (ambient sounds). The research results show that the proposed method is good at overall better speech quality, especially in the case of low SNR values for different color noise and different types of disturbance (ambient sounds). At the same time, it should be noted that the proposed method, apart from the general enhancement of the speech quality, in most cases provides the greatest noise (disturbance) reduction for each of the considered types of noise and types of disturbance, even in the case of a significant contribution of disturbance. Additionally, it is worth noting that the dispersion of the research results for the proposed approach is small in the case of SNR, i.e., the resultant mean value, median value, lower and upper quartile, and the minimum and maximum value excluding outliers, and almost coincides with the adopted axis scale considering the dispersion for each analyzed speech enhancement or noise-reduction methods. On this basis, it can be concluded that the high noise (disturbance) attenuation ability will be maintained with a high probability for various types of other noises (disturbances). In the case of other considered methods, the dispersion is much more noticeable, which indicates significantly changeable effectiveness.

The proposed approach does not require an explicit estimation of the noise level or a priori knowledge of the signal-to-noise ratio as is usually needed in most common speech-enhancement methods. The proposed approach allows for the reconstruction of important features in the time–frequency domain, thanks to which the use of such a block at the beginning of the signal chain in the speech recognition process allows for a significant increase in the accuracy of speech recognition in a situation when the speech signal is noisy or disturbed (e.g., the proposed approach can be used in a radio-illumination decision support system for the captain of a ship in distress situations for better identification and understanding of speech when calling for help and contacting maritime emergency services [58,59]). An effective speech recognition process in difficult enhancement conditions using the proposed approach and deep-learning methods is the future direction of research.

Author Contributions: Conceptualization, P.K.; methodology, P.K.; software, P.K. and W.J.; validation, W.J.; formal analysis, P.K.; investigation, P.K. and W.J.; resources, P.K. and W.J.; data curation, P.K. and W.J.; writing—original draft preparation, P.K.; writing—review and editing, P.K.; visualization, P.K.; supervision, W.J.; project administration, W.J.; funding acquisition, P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by statutory funds of the Faculty of Control, Robotics and Electrical Engineering of the Poznan University of Technology.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to University policy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Recording History. Available online: http://www.recording-history.org (accessed on 1 January 2022).
2. Lindos Electronics. Available online: http://http://www.lindos.co.uk/ (accessed on 1 January 2022).
3. Shapley, G.J. *Sound of Failure: Experimental Electronic Music in Our Post–Digital Era*; University of Technology: Sydney, Australia, 2012.
4. Yang, Y.; SooCho, J.; Lee, B.; Kim, S. A Sound Activity Detector Embedded Low-Power MEMS Microphone Readout Interface for Speech Recognition. In Proceedings of the 2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), Lausanne, Switzerland, 29–31 July 2019 ; pp. 1–6. [CrossRef]
5. Schneider, M. Electromagnetic interference, microphones and cables. *AES J. Audio Eng. Soc.* **2005**, 6339.
6. Gannot, S.; Burshtein, D.; Weinstein, E. Iterative and sequential Kalman filter-based speech enhancement algorithms. *IEEE Trans. Acoust. Speech Signal Process.* **1998**, *6*, 373–385. [CrossRef]

7. Vaseghi, S.V. *Advanced Digital Signal Processing and Noise Reduction*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2008.
8. Kim, J.B.; Lee, K.; Lee, C. On the applications of the interacting multiple model algorithm for enhancing noisy speech. *IEEE Trans. Acoust. Speech Signal Process.* **2000**, *8*, 349–352. [CrossRef]
9. Dendrinos, M.N.; Bakamidis, S.G.; Carayannis, G. Speech enhancement from noise: A regenerative approach. *Speech Commun.* **1991**, *10*, 45–57. [CrossRef]
10. Loizou, P.C. *Speech Enhancement: Theory and Practice*; CRC Press: Boca Raton, FL, USA, 2013.
11. Ephraim, Y.; Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 1109–1121. [CrossRef]
12. Jax, P.; Vary, P. Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong Kong, China, 6–10 April 2003; Volume 1. [CrossRef]
13. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [CrossRef]
14. Kwan, C.; Chu, S.; Yin, J.; Liu, X.; Kruger, M.; Sityar, I. Enhanced speech in noisy multiple speaker environment. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–6 June 2008; pp. 1640–1643 . [CrossRef]
15. Mallat, S. *A Wavelet Tour of Signal Processing*; Academic Press: Cambridge, MA, USA, 2009.
16. Virag, N. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. Acoust. Speech Signal Process.* **1999**, *7*, 126–137. [CrossRef]
17. Sun, C.; Zhu, Q.; Wan, M. A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition. *Speech Commun.* **2014**, *60*, 44–55. [CrossRef]
18. Sun, C.; Xie, J.; Leng, Y. A Signal Subspace Speech Enhancement Approach Based on Joint Low-Rank and Sparse Matrix Decomposition. *Arch. Acoust.* **2016**, *41*, 245–254. [CrossRef]
19. Xian, Y.; Sun, Y.; Wang, W.; Naqvi, S.M. A Multi-Scale Feature Recalibration Network for End-to-End Single Channel Speech Enhancement. *IEEE J. Sel. Top. Signal Process.* **2021**, *15*, 143–155. [CrossRef]
20. Wood, S.U.N.; Rouat, J. Unsupervised Low Latency Speech Enhancement With RT-GCC-NMF. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 332–346. [CrossRef]
21. Chakrabarty, S.; Habets, E.A.P. Time-Frequency Masking Based Online Multi-Channel Speech Enhancement With Convolutional Recurrent Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 787–799. [CrossRef]
22. Lavanya, T.; Nagarajan, T.; Vijayalakshmi, P. Multi-Level Single-Channel Speech Enhancement Using a Unified Framework for Estimating Magnitude and Phase Spectra. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1315–1327. [CrossRef]
23. Tu, Y.H.; Du, J.; Lee, C.H. Speech Enhancement Based on Teacher-Student Deep Learning Using Improved Speech Presence Probability for Noise-Robust Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2080–2091. [CrossRef]
24. Ming, J.; Crookes, D. Speech Enhancement Based on Full-Sentence Correlation and Clean Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 531–543. [CrossRef]
25. Kim, M.; Shin, J.W. Improved Speech Enhancement Considering Speech PSD Uncertainty. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1939–1951 . [CrossRef]
26. Saleem, N.; Khattak, M.I.; Ahmad, S.; Ali, M.Y.; Mohmand, M.I. Machine Learning Approach for Improving the Intelligibility of Noisy Speech. In Proceedings of the 2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 14–18 January 2020; pp. 303–308 . [CrossRef]
27. Choudhury, A.; Roy, P.; Bandyopadhyay, S. Review of Various Machine Learning and Deep Learning Techniques for Audio Visual Automatic Speech Recognition. In Proceedings of the 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC), Taza, Morocco, 26–27 October 2023; pp. 1–10 . [CrossRef]
28. Casey, O.; Dave, R.; Seliya, N.; Sowells Boone, E.R. Machine Learning: Challenges, Limitations, and Compatibility for Audio Restoration Processes. In Proceedings of the 2021 International Conference on Computing, Computational Modelling and Applications (ICCMA), Brest, France, 14–16 July 2021; pp. 27–32 . [CrossRef]
29. Ayhan, B.; Kwan, C. Robust Speaker Identification Algorithms and Results in Noisy Environments. In *Advances in Neural Networks*; Huang, T.; Lv, J.; Sun, C.; Tuzikov, A.V., Eds.; Springer: Cham, Switzerland, 2018; pp. 443–450.
30. Rehr, R.; Gerkmann, T. SNR-Based Features and Diverse Training Data for Robust DNN-Based Speech Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1937–1949 . [CrossRef]
31. Zhang, Q.; Nicolson, A.; Wang, M.; Paliwal, K.K.; Wang, C. DeepMMSE: A Deep Learning Approach to MMSE-Based Noise Power Spectral Density Estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1404–1415 . [CrossRef]
32. Takeuchi, D.; Yatabe, K.; Koizumi, Y.; Oikawa, Y.; Harada, N. Real-Time Speech Enhancement Using Equilibriated RNN. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–9 May 2020; pp. 851–855 . [CrossRef]
33. Zhu, Q.S.; Zhang, J.; Zhang, Z.Q.; Dai, L.R. A Joint Speech Enhancement and Self-Supervised Representation Learning Framework for Noise-Robust Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 1927–1939 . [CrossRef]
34. Shifas, M.P.; Zorila, C.; Stylianou, Y. End-to-End Neural Based Modification of Noisy Speech for Speech-in-Noise Intelligibility Improvement. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2022**, *30*, 162–173 . [CrossRef]

35. Hu, Y.; Li, F.; Li, H.G.; Liu, C. An enhanced empirical wavelet transform for noisy and non-stationary signal processing. *Dig. Signal Process.* **2017**, *60*, 220–229. [CrossRef]
36. Kaiser, J. On a simple algorithm to calculate the 'energy' of a signal. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 3–6 April 1990; Volume 1, pp. 381–384. [CrossRef]
37. Deller, J.; Hansen, J.; Proakis, J. *Discrete-Time Processing of Speech Signals*; Wiley-IEEE Press: Hoboken, NJ, USA, 2000.
38. Rix, A.; Beerends, J.; Hollier, M.; Hekstra, A. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; Volume 2, pp. 749–752. [CrossRef]
39. Gilles, J. Empirical Wavelet Transform. *IEEE Trans. Signal Process.* **2013**, *61*, 3999–4010. [CrossRef]
40. Carvalho, V.R.; Moraes, M.F.; Braga, A.P.; Mendes, E.M. Evaluating five different adaptive decomposition methods for EEG signal seizure detection and classification. *Biomed. Signal Process. Control* **2020**, *62*, 102073. [CrossRef]
41. Donoho, D.; Johnstone, I. Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika* **1994**, *81*, 425–455. [CrossRef]
42. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 5206–5210. [CrossRef]
43. Michael; Li, E.X.D. scivision/Soothing-Sounds: Src/Layout, Black Format, Type Anno. 2021. Available online: https://zenodo.org/record/5574886 (accessed on 1 January 2022).
44. Hu, Y.; Loizou, P. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 334–341. [CrossRef]
45. Scheibler, R.; Bezzam, E.; Dokmanic, I. Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 351–355. [CrossRef]
46. Djahnine, A. Suppression-of-Acoustic-Noise-in-Speech-Using-Spectral-Subtraction. 2021. Available online: https://github.com/AissamDjahnine/Suppression-of-Acoustic-Noise-in-Speech-Using-Spectral-Subtraction- (accessed on 1 January 2022).
47. Bahoura, M.; Rouat, J. Wavelet Speech Enhancement Based on Time-Scale Adaptation. *Speech Commun.* **2006**, *48*, 1620–1637. [CrossRef]
48. Sainburg, T.; Thielk, M.; Gentner, T.Q. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Comput. Biol.* **2020**, *16*, e1008228. [CrossRef]
49. Sainburg, T. timsainb/noisereduce: V1.0. 2019. Available online: https://zenodo.org/record/3243139 (accessed on 1 January 2022). [CrossRef]
50. Yang, Y.; Rao, J. Robust and Efficient Harmonics Denoising in Large Dataset Based on Random SVD and Soft Thresholding. *IEEE Access* **2019**, *7*, 77607–77617. [CrossRef]
51. Yang, Z.X.; Zhong, J.H. A Hybrid EEMD-Based SampEn and SVD for Acoustic Signal Processing and Fault Diagnosis. *Entropy* **2016**, *18*, 112. [CrossRef]
52. PESQ (Perceptual Evaluation of Speech Quality). Wrapper for Python Users. Available online: https://pypi.org/project/pesq/ (accessed on 1 January 2022).
53. Emiru, E.D.; Li, Y.; Xiong, S.; Fesseha, A. Speech Recognition System Based on Deep Neural Network Acoustic Modeling for Low Resourced Language-Amharic. In Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering (ICTCE), Tokyo, Japan, 9–12 November 2019; pp. 141–145. [CrossRef]
54. Cecko, R.; Jamrozy, J.; Jesko, W.; Kusmierek, E.; Lange, M.; Owsianny, M. Automatic Speech Recognition and its Application to Media Monitoring. *Comput. Methods Sci. Technol. CMST* **2021**, *27*, 41–55. [CrossRef]
55. Jesko, W. Vocalization Recognition of People with Profound Intellectual and Multiple Disabilities (PIMD) Using Machine Learning Algorithms. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 2921–2925. [CrossRef]
56. Ravanelli, M.; Parcollet, T.; Plantinga, P.; Rouhe, A.; Cornell, S.; Lugosch, L.; Subakan, C.; Dawalatabad, N.; Heba, A.; Zhong, J.; et al. SpeechBrain: A General-Purpose Speech Toolkit. *arXiv* **2021**, arXiv:2106.04624.
57. Pan, J.; Liu, C.; Wang, Z.; Hu, Y.; Jiang, H. Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMS in acoustic modeling. In Proceedings of the 2012 8th International Symposium on Chinese Spoken Language Processing, Hong Kong, China, 5–8 December 2012; pp. 301–305. [CrossRef]
58. Banaszek, A.; Lisaj, A. The Concept of Intelligent Radiocommunication System for Support Decision of Yacht Captains in distress situations with use of neural network computer systems. *Procedia Comput. Sci.* **2022**, *207*, 398–407. [CrossRef]
59. Banaszek, A.; Lisaj, A. Advanced methodology for multi-way transmission of ship data treatment from mechanical-navigational technical state sensors with using computational neural network computer systems. *Procedia Comput. Sci.* **2022**, *207*, 388–397. [CrossRef]