Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 6582296, 11 pages
https://doi.org/10.1155/2023/6582296

Research Article

Adaptive Noise Reduction Algorithm Based on SPP and NMF for Environmental Sound Event Recognition under Low-SNR Conditions

Li-Yan Luo,[1,2] Shu-Ting Guo,[2] Mei Wang,[1,3] Hong-Bing Qiu,[1,2] and Zheng-Hong Liu[1,2]

[1] *Provincial Ministry of Education Key Laboratory of Cognitive Radio and Signal Processing, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China*
[2] *Guangxi Key Laboratory of Wireless Broadband Communication and Signal Processing, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China*
[3] *College of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi 541007, China*

Correspondence should be addressed to Mei Wang; 414863983@qq.com

Received 10 March 2022; Revised 21 November 2022; Accepted 12 December 2022; Published 21 January 2023

Academic Editor: Petros Nicopolitidis

Nonnegative matrix factorization- (NMF-) based noise reduction methods can effectively improve the performance of environmental sound recognition. However, when the environmental sound overlaps highly with the noise, the spectral line loss and noise residue will occur in the low signal-to-noise ratio (SNR) condition. An adaptive noise reduction algorithm was proposed in this paper. First, noisy environmental sound is separated into estimated noise and environmental sound using NMF. Then, the estimated noise is used to calculate sound presence probability (SPP), which is adapted to decrease spectral line loss and achieve accurately estimated noise. Subsequently, the estimated noise combines with noisy environmental sound to obtain the estimated environmental sound. Finally, SPP is applied to reduce residual noise in the estimated environmental sound and reconstruct the environmental sound. The simulation results demonstrate that the proposed algorithm outperforms the traditional algorithms and NMF-based methods in terms of perceptual evaluation of speech quality (PESQ) and global SNR with increase of X% and X%, respectively. Moreover, the proposed method can effectively improve the environmental sound recognition rate. Particularly, the proposed method makes a 16.2% increase of F1-score in car horn recognition under the realistic acoustic condition.

1. Introduction

The recognition of environmental sound (ES) enables the monitoring of certain specific event, since ES have the potential of characterizing the surrounding environment. However, one key factor that affects detection and classification performance is the diverse and unpredictable interference noise in the real-life scenarios [1]. Therefore, noise reduction (NR), as a part of the preprocessing of ES, has important application prospects in human-computer interaction [2], animal behavior monitoring [3], anomalous sounds for machine condition monitoring [4], and domestic risk scenarios [5]. Speech as the first studied ES, some representative NR methods such as spectral subtraction (SS) [6], Wiener filtering (WF) [7], minimum mean square error (MMSE) [8, 9], and short-time spectral amplitude (STSA) [10] have been proposed. By the virtue of highly flexible and ease of implementation, these methods are applied for nonspeech NR in WASNs [11, 12]. It is important to notice that these works have limited performance to suppress nonstationary noise despite the various contributions, since their strong assumption is based on stationary noise, especially under low-SNR conditions [13].

To cope with the nonstationary noise, noise estimation is performed before the NR gain function, which provides accurate noise spectrum to evaluate the WF [14] or to estimate the a priori SNR in the MMSE [15]. In [16], the author used improved minimal controlled recursive averaging to estimate noise and combined with used optimally modified log-spectral amplitude (OMLSA) to obtain promising denoised performance. Although such approaches can estimate noise spectrum continuously during ES activity, it does not respond well to the increasing noise levels at low-SNR conditions, which leads to the underestimation of the noise spectrum and annoying residual noise.

In order to provide consistent performance in highly nonstationary noisy environments under low-SNR conditions, NMF models the input noisy as a weighted sum of the nonnegative basis from clean ES and noise [17]. Since the same acoustic elements exit for ES and noise, NMF is not effective in separating regions where the ES and noise spectral bands are heavily mixed [17]. Subsequent works attempt to optimize separation rules [18, 19] and postprocessing [20, 21] to improve the quality of separated ES. For instance, a weighted NMF for interpolating missing data is presented [18], which can address the overestimate of the values in the masked regions and the computational cost equivalent to standard NMF. In [20], Lee et al. introduced spectral-temporal speech presence probabilities (SPP) to reconstruct the regions of the separated speech with severe spectral leakage to suppress the residual noise components. As these methods are tempted to process the separated ES as denoised result, they ignore the key issue that the spectral line loss of separated ES is beyond repair, which affects the clarity of denoised ES.

Benefiting from the high learning capacity of deep neural network (DNN), a hybrid model which combined DNN and NMF is proposed. In [22], DNN is applied to initialize the activation matrix of NMF, with which the performance is slightly better than the traditional NMF. Another approach estimates the activation matrix through a DNN and then reconstructs ES through multiplying it with the basis matrix [23]. However, DNN-based methods require enormous number of clean ESs which is difficult to obtain in advance, such as gunshots and explosions. Without a sufficient training dataset, these methods might be overfitted.

Furthermore, NMF is considered to be sensitive to nonstationary noise and low-SNR conditions when using the separated noise as the noise estimation result [22]. Based on this, sparse and low-rank NMF with Kullback-Leibler divergence is presented to noise estimation [24]. Lai et al. [25] applied NMF for noise estimation in combination with the Wiener filter gain function to obtain enhanced speech with high quality and intelligibility under challenging conditions. However, the above methods rely on a supervised learning approach and still fail to improve the robustness of the algorithm in an unseen noise environment.

Moreover, current NR algorithms for monitoring system draw on speech enhancement schemes directly. In automatic recognition system of porcine abnormalities [26], SS is employed to suppress the pigpen background noise to improve the detection performance. In [4], researchers use the noisereduce libraries of Python 3.7 for removing background noise in the cattle farm and the analog white noise of the microphone. After removing the noise, the performance of cattle vocal classification improves from 91.38% to 94.18%. Such methods lack consideration of monitoring sound characteristics; for instance, Xu et al. [27] employed the improved control recursive averaging algorithm to estimate noise which would be disabled when the ES of nonspeech changes more slowly than noise.

The contributions of this paper are summarized as follows: (1) an adaptive NR algorithm using the semisupervised learning mode, which reduces the interference from segments of non-ES and further improves the performance of recognition. (2) SPP-based threshold determination is presented for locating the frame where ES is vocalized in the sound clip. (3) To verify the validity of the proposed algorithm in the monitoring system, experiments of simulations and realistic acoustic conditions were conducted on the nonspeech datasets. The experimental results showed that the proposed algorithm achieves good NR performance.

The paper is organized as follows. Section 2 introduces the NMF-based NR technique. Section 3 presents the framework of the proposed algorithm and details the algorithm. Section 4 analyzes the experimental results and evaluates the performance of the proposed algorithm. Finally, we draw the conclusions in Section 4.

2. The NMF-Based NR Technique

NMF is a technique of source separation to additive mixtures by using the basis matrix. Since NMF is capable of interpreting the local properties of the image, NMF is proposed as face recognition technology initially [28]. Recently, NMF has been studied for blind source separation [29, 30] and NR, on account of sound that can be converted to the spectrogram form.

Consider the representation of a noisy ES $y(n)$ in the time-frequency (T-F) domain as the sum of environmental $X(f,t)$ sound and noise $N(f,t)$,

$$Y(f,t)=X(f,t)+N(f,t), \tag{1}$$

where f is the frequency bins and t is the time index.

To satisfy the nonnegativity constraint of NMF in the input matrix, the nonnegative real-valued matrix of $f \times t$ dimensions, namely, the T-F amplitude spectrum matrix $|Y(f,t)|$ of noisy ES, is taken as the decomposition matrix. In the decomposition of the matrix, the nonnegative basis matrix $W \in \mathrm{R}^{f \times t}$ and the activation matrix $W \in \mathrm{R}^{f \times t}$ are such that $|Y(f,t)| \approx \mathrm{WH}$. NMF allows the original high-dimensional matrix to be approximately decomposed into the multiplication form of a low-rank matrix. The potential structure of the original structure is captured by W, and H is its corresponding T-F gain. Therefore, the rank k is required to be much smaller than f or t and satisfies $k < \mathrm{ft}/(f+t)$.

To find optimize W and H, the Euclidean distance is used to quantify the approximate mass of the decomposition to minimize the reconstruction error,

$$\mathrm{D}(W,H)=\|Y-\mathrm{WH}\|_{\mathrm{F}}^{2}. \tag{2}$$

Then, the W and H iterative multiplication update rules are as follows [31]:

$$W^i = W^{i-1} \otimes \frac{Y \cdot \left(H^{i-1}\right)'}{W^{i-1} \cdot H^{i-1} \cdot \left(H^{i-1}\right)'}, \tag{3}$$

$$H^i = H^{i-1} \otimes \frac{Y \cdot \left(W^{i-1}\right)'}{\left(W^{i-1}\right)' \cdot W^{i-1} \cdot H^{i-1}}, \tag{4}$$

where i is the number of iterations and $\otimes$ represents the multiplication of the corresponding elements of the matrix.

For NR tasks, W can be rewritten as a joint dictionary form of ES and noise, namely, $W = [W_X\, W_N]$. Similarly, $H = [H_X\, H_N]$.

Accordingly, the decomposition process of the noisy T-F amplitude spectrum matrix is shown in Figure 1. By matrix operation, the T-F amplitude spectrum matrix of ES $|X(f, t)|$ and noise $|N(f, t)|$ can be approximated as

$$\begin{aligned} |X(f, t)| &\approx W_X H_X, \\ |N(f, t)| &\pounds W_N H_N. \end{aligned} \tag{5}$$

$|X(f, t)|$ can be considered as denoised output.

When NMF decomposing the ES matrix, the basis matrix is activated unstably because the same sound elements overlap between the basis matrices of different sounds [24]. Thus, there is a mutual spectral component leakage between the separated $|X(f, t)|$ and $|N(f, t)|$ in the strong noise area where the spectrum of ES and noise overlaps highly. This leads to the spectral line loss of the separated ES which affects the quality and clarity of the denoised output. Therefore, a reconstruction scheme based on separated noise is proposed for the separated ES distortion region.

3. The Proposed Denoised Scheme

The block diagram of the proposed algorithm scheme is shown in Figure 2. An unsupervised learning method was applied to the captured noise by building a noise buffer, and the separated noise from NMF was considered as the preliminary noise estimation result. In the noise processing stage, an SPP algorithm was used to adaptively suppress the leaked ES component in the separated noise. Combined with an OMLSA spectral estimator, the high-frequency structural information of ES can be retained to further improve the ES quality. Finally, the ES output can be enhanced through a residual noise suppression process. The key steps include building noise buffer, adaptive noise processing, and residual noise suppression.

3.1. Semisupervised NMF. For the monitoring system, the target ES to be monitored is clear. Thus, W_X is prefetched from the ES dataset. Without presetting the type of noise in the environment, we extract W_N from building noise buffer and combine W_N with the pretrained W_X to achieve semisupervision NMF. The noise buffer is considerable to capture noise-only segments to update the noise basis matrix online, which meets the demand of NR in unknown noise environment.

In the test stage, unsupervised NMF [28] is executed. First, initialize W_X, W_N, H_X, and H_N with negative random values. Then, W_N is updated iteratively by Equation (3), and H_X and H_N are updated iteratively by Equation (4). Iteration stops when $i = i_{\max}$ or the convergence Equation (2) is satisfied, resulting in the separated T-F amplitude spectrum matrix of ES $|\widehat{X}_1(f, t)|$ and the separated noise T-F amplitude spectrum matrix $|\widehat{N}_1(f, t)|$.

3.2. Adaptive Noise Suppression. Due to the overlap of acoustic elements, semisupervised NMF is only used as a preliminary separation. The separation results are used to calculate instantaneous SNR and the presence probability of ES in the separation noise, providing accurate prior information for the subsequent estimation of clean ES. In this stage, SPP is used to adaptively suppress the leaked ES component in the separated noise to improve the performance of the OMLSA estimator.

First, it is assumed that the T-F amplitude spectrum of ES and noise satisfies the complex Gaussian distribution and that $H_1(f, t)$ and $H_0(f, t)$, respectively, represent the presence and absence of ES at the T-F point. Then, the conditional probability distribution function of the observed signal can be given by the variance of ES $\sigma_X\,(f, t)$ and noise $\sigma_N\,(f, t)$.

$$\begin{aligned} p(Y(f, t)|H_0(f, t)) &= \frac{1}{\pi\sigma_X(f, t)} \exp\left\{-\frac{|Y(f, t)|^2}{\sigma_N(f, t)}\right\}, \\ p(Y(f, t)|H_1(f, t)) &= \frac{1}{\pi(\sigma_X(f, t) + \sigma_N(f, t))} \\ &\quad \cdot \exp\left\{-\frac{|Y(f, t)|^2}{\sigma_X(f, t) + \sigma_N(f, t)}\right\}. \end{aligned} \tag{6}$$

By applying Bayes' rule, the conditional probability of ES presence $p(f, t)$ is given by the following formula [10]:

$$p(f, t) = \left\{1 + \frac{q(f, t)}{1 - q(f, t)}(1 + \xi(f, t)) \exp\left(-\upsilon(f, t)\right)\right\}^{-1}, \tag{7}$$

where $q(f, t) \triangleq P(H_0(f, t))$ is the prior probability of ES absence and $\upsilon(f, t)$ is obtained from the ratio of prior and posterior SNR, namely,

$$\upsilon(f, t) \triangleq \frac{\gamma(f, t)\xi(f, t)}{1 + \xi(f, t)}. \tag{8}$$

Since the T-F points of mutual leakage between the ES and noise are the regions with serious spectrum aliasing, instantaneous SNR is considered to determine the T-F

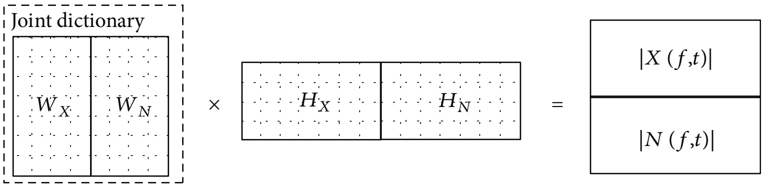


Figure 1: NMF for noise reduction.

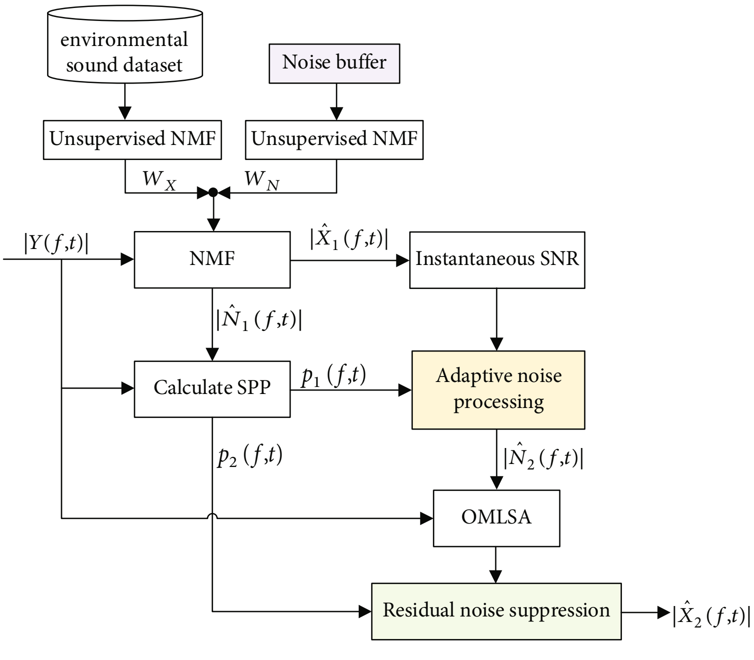


Figure 2: Block diagram of the proposed algorithm.

points (t_i, f_i) that need to be reconstructed in the separated noise.

$$(t_i, f_i) = 10 \log \frac{\left|\widehat{X}_1(t,f)\right|^2}{\left|\widehat{N}_1(t,f)\right|^2} < \text{inputSNR}(t,f). \qquad (9)$$

According to Equation (7), if the conditional probability of ES presence in the separated noise $N_1(f,t)$ is $p_1(f,t)$ and $p_1(f,t) \in [0,1]$, then

$$\begin{aligned} N_2(f_2, t_i) &= (1 - p_1(f_i, t_i)) \times N_1(f_i, t_i), \\ N_3(f_i, t_i) &= \alpha(f_i, t_i) N_1(f_i, t_i) + (1 - \alpha(f_i, t_i)) N_2(f_i, t_i), \end{aligned} \qquad (10)$$

where $\alpha(t_i, f_i)$ is an adaptive factor that automatically adjusts the noise component of the separated noise according to the instantaneous SNR level. That is, $\alpha(t_i, f_i)$ is a weighting factor of the power ratio of the ES component to the noise component in the aliasing region of the reaction spectrum. At higher SNRs, a larger α should be used to avoid weakening of the separated noise components, resulting in more residual noise in the denoising ES. At low-SNR conditions, a smaller α should be used to avoid denoising ES distortion.

$$\alpha(f,t) = \begin{cases} 0.95, & p_1(f,t) = 1, \\ \dfrac{1}{e^{-\text{instantaneous SNR}(t,f)}}, & 0.2 < p_1(f,t) < 1, \\ 0.05, & p_1(f,t) \le 0.2. \end{cases} \qquad (11)$$

Then, the separated noise $N_3(f,t)$, which is processed by adaptively suppressing the ES component, is used as the input for the OMLSA estimator to provide noise information in the noisy ES, to accurately determine the spectral gain function $G(f,t)$ [10].

$$G(f,t) = \left\{G_{H_1}(f,t)\right\}^{p(f,t)} \cdot G_{\min}^{1-p(f,t)}. \qquad (12)$$

$G_{\min}(f,t)$ is the lower threshold of $G(f,t)$ for the absence of ES, while $G_{H1}(f,t)$ is the spectral gain function that minimizes the logarithmic MMSE of the estimated ES amplitude and the true amplitude as an optimization objective in the presence of ES.

$$G_{H_1}(f,t) = \frac{\xi(f,t)}{1+\xi(f,t)} \exp\left(\frac{1}{2}\int_{v(f,t)}^{\infty} \frac{e^{-t}}{t}dt\right). \tag{13}$$

As the noisy ES had been separated by NMF, $\sigma_X(f,t)$ and $\sigma_N(f,t)$ can be given by $\widehat{X}_1(f,t)$ and $\widehat{N}_1(f,t)$, respectively, when computing $\gamma(f,t)$ and $\xi(f,t)$ using a decision-direction approach.

$$\begin{aligned}\gamma(f,t) &\triangleq \frac{|Y(f,t)|^2}{\sigma_N(f,t)} = \frac{|Y(f,t)|^2}{\left|\widehat{N}_1(f,t)\right|^2},\\ \xi(f,t) &\triangleq \frac{|X(f,t)|^2}{\sigma_X(f,t)}\\ &= aG_{H_1}^2(f,t-1)\gamma(f,t-1)\\ &\quad + (1-a)\max[\gamma(f,t)-1,0].\end{aligned} \tag{14}$$

Finally, by applying the OMLSA spectral gain function to $|Y(f,t)|$, the T-F amplitude spectrum of the denoised ES can be obtained as

$$\left|\widehat{X}_2(f,t)\right| = G(f,t)|Y(f,t)|. \tag{15}$$

3.3. Residual Noise Suppression. If the ES type is speech, the $|\widehat{X}_2(f,t)|$ obtained from Equation (15) is used as the final denoised output. However, for NR in nonspeech monitoring system, the input data is a sound segment of about five seconds obtained by endpoint detection processing. The presence of non-ES frames in the input data, especially those with residual noise, would affect recognition performance. In this stage, we design the T-F weighting factor $\beta(f,t)$ to further enhance the $|\widehat{X}_2(f,t)|$.

$\beta(f,t)$ extracts the frame with the highest probability of ES, while suppressing the frames dominated by interference noise.

$$\beta(f,t_j) = \begin{cases} 1, & \frac{1}{\text{wlen}}\sum_{\text{k}=1}^{\text{wlen}} A(f_k) \geq \delta_2,\\ 0, & \text{otherwise,}\end{cases} \tag{16}$$

where $A(f)_{\text{wlen}\times 1}$ and δ_2 represent the flag matrix and the decision threshold, respectively, to determine the presence of ES in each frame. By comparing $p_2(f,t)$ calculated from $|\widehat{X}_2(f,t)|$ with the decision threshold of the frequency point δ_2, $A(f)_{\text{wlen}\times 1}$ is given by

$$A(f_k) = \begin{cases} 1, & p_2(f_k,t_j) \geq \delta_1,\\ 0, & \text{otherwise.}\end{cases} \tag{17}$$

Therefore, the final output of the denoised nonspeech can be expressed as

$$\left|\widehat{X}_3(f,t)\right| = \beta(f,t)\left|\widehat{X}_2(f,t)\right|. \tag{18}$$

4. Experimental Results

This study evaluated the denoised performance of the proposed adaptive NR method and compared it with the standard NMF [28], RNMF [20], and the traditional speech enhancement methods Wiener [8], STSA [10], and OMLSA [32] on the ES datasets. In addition, we apply the proposed method to the nonspeech monitoring system and conducted simulations and real experiments.

4.1. Datasets and Experimental Parameter Setting. ES datasets consist of the Google dataset [33] and TIMIT dataset [34], including car horn, scream, gunshot, and speech. The dataset is divided according to the ratio of 7 : 3 as training set and test set, respectively. Noise test set including nonstationary noise babble, factory2, F16, destroyerops, pink, and white were selected from NOISEX-92 dataset [35], and the natural environment noises rain and wind were selected from ESC-50 database [36]. The SNRs ranged in volume from −5 to 5 dB, and noise that did not overlap with the training set was added to the clean sound of the test set to generate enough noisy ES to evaluate the performance of the proposed algorithm. All audio was resampled to 16 kHz, and the time domain signal was converted into a T-F amplitude spectrum by STFT, with a hamming window length of 32 ms and a 50% frame shift. The dimensions of the ES and noise basis matrices were derived from experience and set as 90 and 60, respectively.

4.2. Evaluation Metrics. In the speech-based noise suppression experiments, global SNR and PESQ [37] were selected as the evaluation metrics. Global SNR is defined as the power ratio between ES and noise over all T-F regions, reflecting the relative magnitude of the two. Through the improvement of global SNR, the amount of noise rejection of the algorithm from an objective perspective is given for speech quality evaluation. PESQ is considered to be an objective expression of subjective evaluation, which can compensate for the lack of global SNR measurement. In nonspeech experiments, the F-score index of the sound event recognition model was used to reflect the noise suppression ability of the algorithm [38]. This index is the result of balancing the precision and recall indexes comprehensively, and the index is defined as follows:

$$\begin{aligned}P &= \frac{\text{TP}}{\text{TP}+\text{FN}},\\ R &= \frac{\text{TP}}{\text{TP}+\text{FN}},\\ \text{F-score} &= \left(1+\beta^2\right)\frac{P\cdot R}{\beta^2(P+R)},\end{aligned} \tag{19}$$

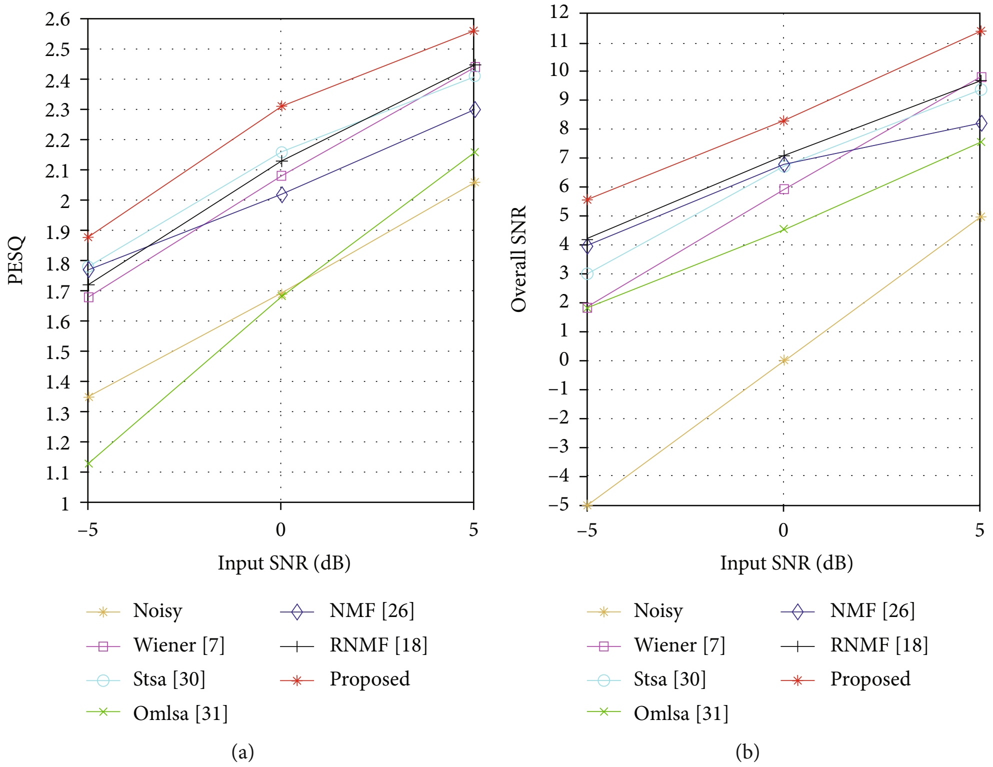


Figure 3: Performance of the proposed algorithm compared with other algorithms on the speech dataset. (a) Global SNR; (b)PESQ.

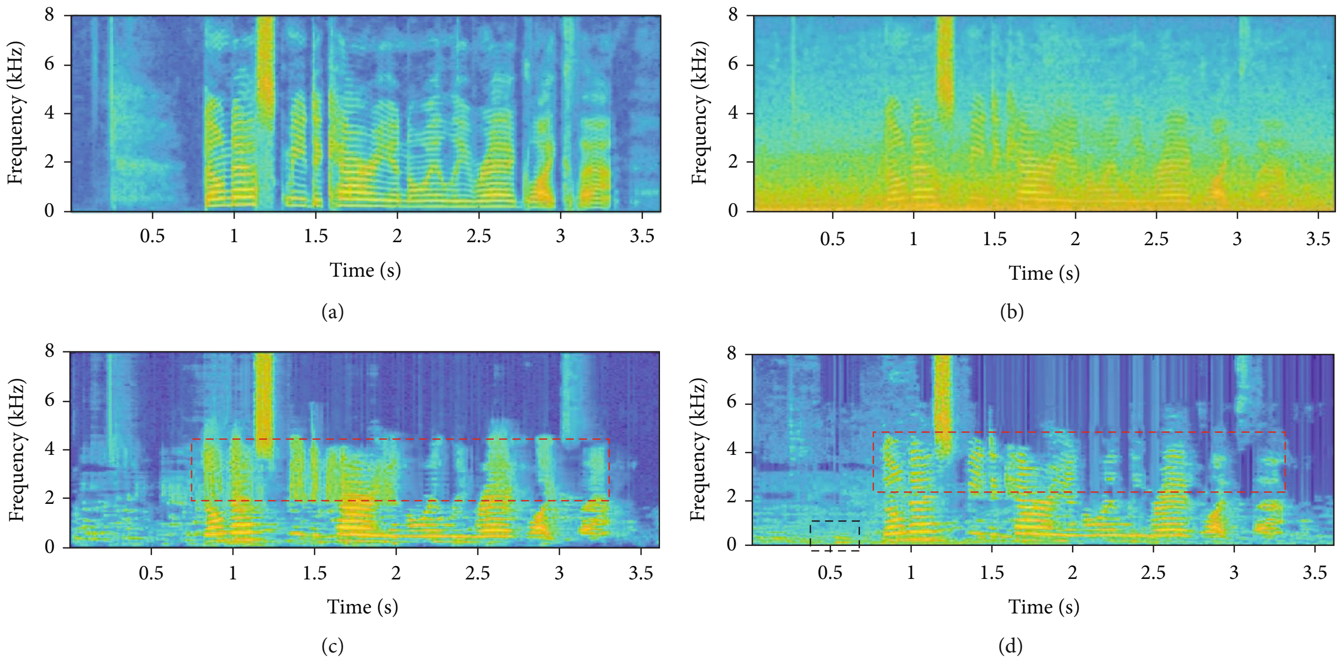
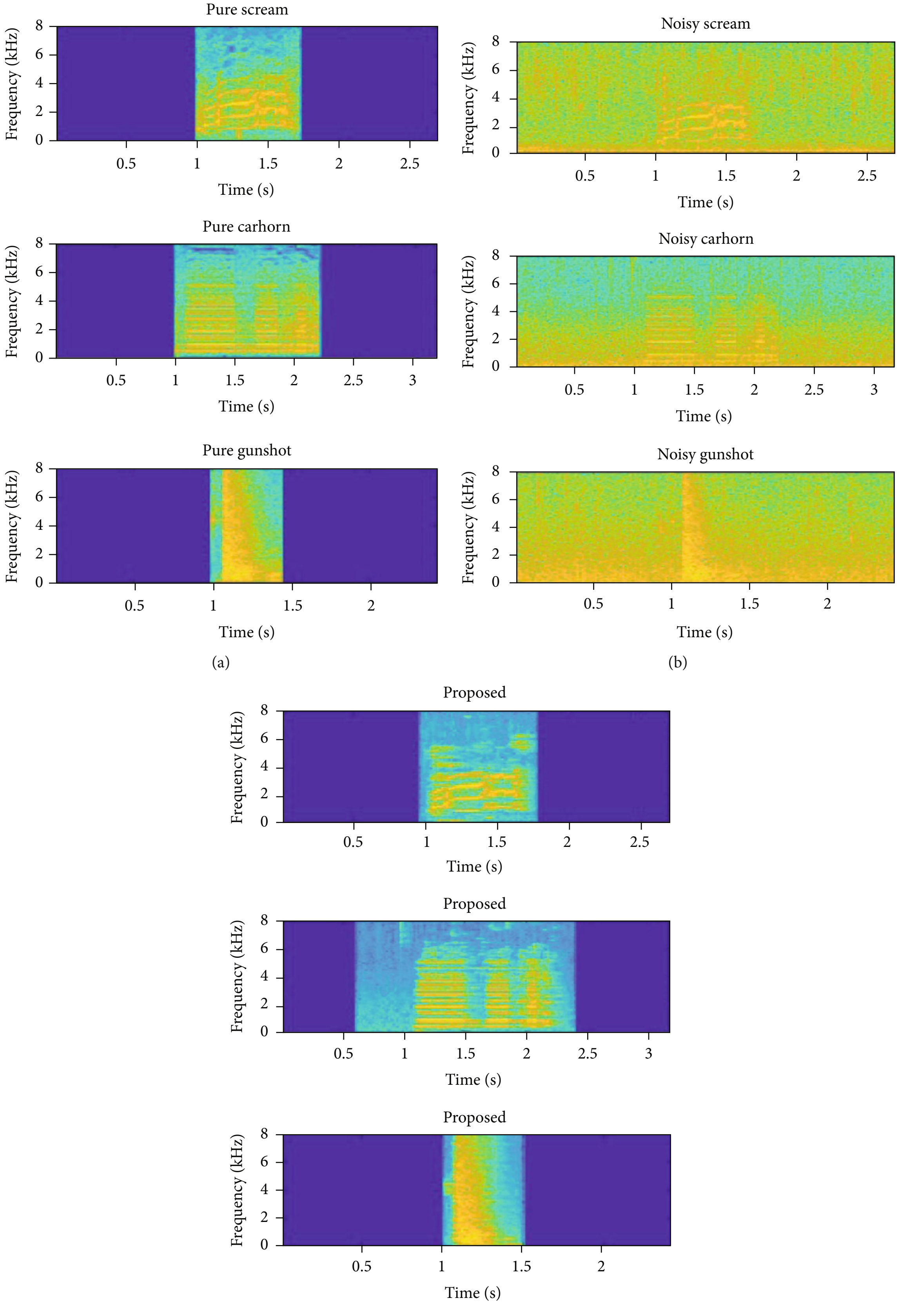


Figure 4: T-F spectrogram. (a) Clean ES; (b) noisy ES; (c) RNMF [16]; (d) the proposed algorithm.

where the TP, TN, FP, and FN codes are positive, negative, false positive, and false negative, respectively [39]. P is the precision rate, while R is the recall rate. In this experiment, when $\beta = 1$, the weight of precision and recall rates was the same; that is, the F1-score index was selected.

In addition, since T-F spectrogram analysis is a common method for analyzing the frequency and level of time-varying signals, this paper also uses a T-F spectrogram as an evaluation metrics in both speech and nonspeech experiments.



(a)

(b)

(c)

Figure 5: T-F spectrogram. (a) Clean ES; (b) noisy ES at SNR = 0 dB; (c) the proposed algorithm.

4.3. Performance Comparison of Different Algorithms to Suppress Nonstationary Noise under Low-SNR Conditions. To directly reflect the nonstationary noise suppression ability of the proposed algorithm at low-SNR conditions, noisy speech, which was disrupted by six nonstationary noises from NOISEX-92 database, was selected for the experiments. The proposed algorithm was compared with a traditional speech enhancement algorithm [7, 10, 32], standard NMF [28], and reconstructed NMF (RNMF) [20].

According to Figure 3, the proposed algorithm is superior to the compared algorithm in both global SNR and PESQ metrics. When compared with other algorithms under the three input SNR conditions, the proposed algorithm improved by 0.13–0.60 (0.26 on average) under the PESQ index. With the global SNR indicator, the minimum improvement was 1.44 dB, the maximum improvement was 3.78 dB, and the average improvement was 2.39 dB. As shown in Figure 3(a), when the input SNR was equal to −5 dB or 0 dB, NMF, RNMF, and the proposed algorithm were superior to the traditional algorithm under the global SNR index. This showed the superiority of the algorithm based on NMF under low-SNR conditions and its ability to better protect speech quality. At SNR = 5 dB, NMF was lower than the conventional Wiener and STSA algorithms, which is consistent with the previous conclusion that noise spectrum leakage and high spectral line blurring in separated ESs would result in poor noise suppression. The conclusion is also reflected in Figure 3(b), where the PESQ values of NMF were lower than that of the Wiener or STSA algorithms at both 0 dB and 5 dB.

Comparing the T-F spectrograms found in Figures 4(c) and 4(d), the enhanced speech of the proposed algorithm had clear spectral lines and less residual noise in the overlapping region of speech and noise. This shows that the proposed algorithm can effectively suppress the spectral components leaked during NMF separation while preserving the high-frequency information of speech (indicated in the red-boxed region) and thus had the highest global SNR and PESQ performance. Although RNMF addressed the noise spectrum leakage of NMF through reconstructing, there is still a loss of high-frequency information (indicated in the red-boxed region), because the reconstruction of RNMF is based on the separation of ES. This is also why RNMF exhibits limited noise rejection performance (Figure 3). In the 0–0.7 s region for the T–F spectrograms, fewer isolated noise fragments remained after the processing of the proposed algorithm, indicating that the proposed algorithm suppresses nonstationary noise more thoroughly.

4.4. Influence of Environmental Sound Noise Reduction on the Recognition Result. NR is an indispensable part of recognition task preprocessing; this paper selected three types of nonspeech, namely, car horns, screams, and gunshots, as the recognized sound events for the experiments. The noise suppression performance of the proposed algorithm can be measured by the improvement of model recognition ability before and after NR. Nonstationary noises, such as rain, wind, and babble, were selected as interference noises in the outdoor environment. Noisy ES under low-SNR conditions was synthesized for the experiments. The recognition model is based on the two-input convolutional neural network of the previous work [38].

Table 1: F1-score for the recognition task.

Input SNR	Method	Scream	Car horn	Gunshot	Average
5 dB	Noisy	87.9	76.2	77.3	80.5
	Pro	90.0	92.8	92.0	**91.6**
0 dB	Noisy	82.0	75.7	75.0	77.6
	Pro	88.0	91.2	89.1	**89.4**
-5 dB	Noisy	67.2	65.5	67.5	66.7
	Pro	80.8	85.7	84.0	**83.5**
Average	Noisy	79.0	72.4	73.3	**74.9**
	Pro	**86.2**	**89.9**	**88.4**	**88.2**

Figure 5 shows the T-F spectrogram of the nonspeech used as the input for the recognition model, displaying three kinds of nonspeech ES destroyed by nonstationary rain noise at 0 dB SNR condition. Comparing Figures 5(b) and 5(c), the proposed method can locate the segments with ES and set T-F weights of non-ES segments to 0 (the dark blue part). This demonstrated that the T-F weighting factor we designed is effective. Figure 4(d) shows the NR results without using the T-F weighting factor, and we can see that there is noise residue in the low frequency at 0.5 seconds (indicated in the black-boxed region). Setting zero to the weights of non-ES segments can solve the performance degradation caused by residual noise of pure noise segments. Thorough suppression of residual noise is more conducive to subsequent ES recognition. In addition, by comparing Figures 5(a) and 5(c), the spectrum structure of the nonspeech reconstructed from the proposed method processing was mainly intact. In the low-frequency range where noise damage was serious and in the high-frequency range where nonstationary noise is the main component, there was less noise residue.

The F1-score of the noisy ES and the denoised ES is listed in Table 1. According to the table data, the ES processed by the algorithm in this paper had higher F1-score. On average, F1-score improved by 11.1%, 11.8%, and 16.8%, respectively, under the three low-SNR conditions. At -5 dB, the largest improvement in the model indicates the effectiveness of the proposed algorithm at low-SNR conditions and can significantly improve the recognition performance of the recognition model.

4.5. Analysis of the Application of the Proposed Algorithm in Environmental Sound Recognition Systems. To verify the effectiveness of the proposed algorithm in improving the performance of the recognition systems under real scenarios, the recognition system designed in the previous work [40, 41] was used in this paper for real-time environmental sound data acquisition and recognition processing. To ensure the authenticity of the experiment, a car horn was collected from the Guilin University of Electronic and Technology. Along with a car horn used by electric vehicles, two mobile-side devices and two fixed-side devices were placed for real-time data collection. The collection scenario

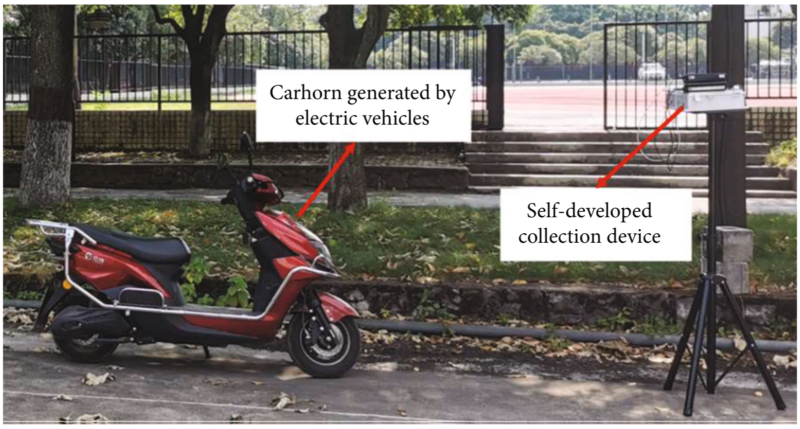
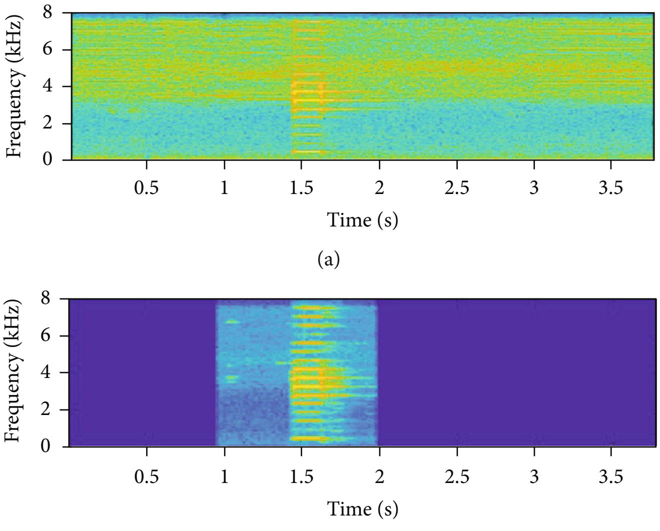

Carhorn generated by electric vehicles
Self-developed collection device

Figure 6: Collection scenarios and device placement.



(a)

(b)

Figure 7: T-F spectrogram of real ES. (a) Noisy ES; (b) the proposed algorithm.

and device placement are shown in Figure 6. The experiment was conducted in July of 2021, and the main background sounds on the campus were cicadas, birds, wind, and building noises. The energy range of the noise measured with a sound level meter at the acquisition microphone was 70 to 90 dB.

The duration of each piece of sound data collected in this experiment was 3-5 s, each with 16 kHz sampling rate, and saved in WAV format. Figure 7 shows an example of the captured noisy ES and ES processed by the proposed algorithm. As shown in Figure 7(b), in the frequency band of 4-8 kHz with cicadas, the spectrum line of the denoised car horn was clear. At the same time, the noise was completely muted or set to a smaller amplitude in the pure noise segments. This demonstrates the ability of the proposed algorithm to learn the noise feature in the real system, successfully extracting the nonspeech components from the noisy bands. By suppressing the noise components, the proposed algorithm can prevent the low recognition rate caused by interference noise in the acquired ES.

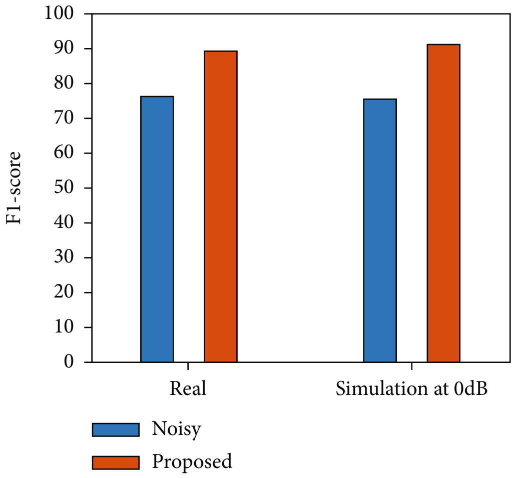


Figure 8: F1-score of the real experiment and the simulation experiment at 0 dB.

In this experiment, the F1-score in Experiment 2 was used to measure the NR performance of the proposed algorithm. Figure 8 shows that in a real environment, the F1-score of the model for handling the noisy and denoised car horn is 76.2% and 88.8%, respectively. The 16.2% improvement in the F1-score after NR was also consistent with the 15.5% improvement at $\mathrm{SNR} = 0\,\mathrm{dB}$ demonstrated in the simulation experiment of Experiment 2.

5. Conclusion

In this paper, an adaptive NR algorithm used in various noisy environments was proposed. First, a noise buffer was set to implement semisupervised NMF so that the algorithm can suppress the unseen noise to improve algorithm robustness. Next, the proposed algorithm designed an adaptive weight factor based on SPP to suppress leaked ES components in the separation noise, avoiding misclassification of the ES as noise for suppression and solving the distortion problem. In addition, to reduce the residual noise interference of non-ES segments on recognition, the T-F threshold was used for each frame of the OMLSA estimator output. The results show that the proposed algorithm outperformed other methods in terms of the average PESQ and SNR. Under realistic acoustic conditions, the proposed algorithm combined with the monitoring system significantly improved the recognition performance of the monitoring system.

However, the ES and basis matrices are in one-to-one correspondence. Therefore, in the system with high real-time performance, it is necessary to increase the calculation cost to determine the corresponding basis matrix. In future works, the optimization of the algorithm will be considered to improve its computational speed and suitability for tasks where real-time performance is required.

Data Availability

Data are available at Google AudioSet: https://ieeexplore.ieee.org/document/7952261.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62201163), the Project of Guangxi Natural Science Foundation (No. 2020GXNSFAA159004), the Project of Guangxi Technology Base and Talent Special Project (No. GuiKe AD20159018), and the Fund of Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education (No. CRKL200104).

References

[1] J. Y. Son and J. H. Chang, "Attention-based joint training of noise suppression and sound event detection for noise-robust classification," *Sensors*, vol. 21, no. 20, p. 6718, 2021.

[2] A. R. Avila, Z. Akhtar, J. F. Santos, D. OShaughnessy, and T. H. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 177–188, 2021.

[3] D. H. Jung, N. Y. Kim, S. H. Moon et al., "Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering," *Animals*, vol. 11, no. 2, p. 357, 2021.

[4] J. A. Lopez, G. Stemmer, P. Lopez-Meyer, P. Singh, J. A. del Hoyo Ontiveros, and H. A. Cordourier, "Ensemble of complementary anomaly detectors under domain shifted conditions," in *Detection and Classification of Acoustic Scenes and Events*, pp. 11–15, Intel Corporation, 2021.

[5] A. García-Domínguez, C. E. Galván-Tejada, R. F. Brena et al., "Children's activity classification for domestic risk scenarios using environmental sound and a Bayesian network," *Healthcare*, vol. 9, no. 7, p. 884, 2021.

[6] Y. Ephraim and H. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[7] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, pp. 629–632, 1996.

[8] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[9] B. M. Mahmmod, S. H. Abdulhussain, M. A. Naser, M. Alsabah, and J. Mustafina, "Speech enhancement algorithm based on a hybrid estimator," *IOP Conference Series: Materials Science and Engineering*, vol. 1090, no. 1, 2021.

[10] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, 2005.

[11] F. D. Arce, M. Moonen, M. Verhelst, and A. Bertrand, "Adaptive quantization for multichannel Wiener filter-based speech enhancement in wireless acoustic sensor networks," *Wireless Communication and Mobile Computing*, vol. 2017, article 3173196, 15 pages, 2017.

[12] J. Vandendriessche, N. Wouters, B. da Silva, M. Lamrini, M. Y. Chkouri, and A. Touhafi, "Environmental sound recognition on embedded systems: from FPGAs to TPUs," *Electronics*, vol. 10, no. 21, p. 2622, 2021.

[13] Itu, *Transmission Systems and Media, Digital Systems and Networks-Voice Enhancement Devices*, ITU-T Recommendation G.160, 2008.

[14] N. Saleem, M. I. Khattak, G. Witjaksono, and G. Ahmad, "Variance based time-frequency mask estimation for unsupervised speech enhancement," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 31867–31891, 2019.

[15] Q. Q. Zhang, M. J. Wang, Y. Lu, L. Zhang, and M. Idrees, "A novel fast nonstationary noise tracking approach based on MMSE spectral power estimator," *Digital Signal Processing*, vol. 88, pp. 41–52, 2019.

[16] T. D. Tran, Q. C. Nguyen, and D. K. Nguyen, "Speech enhancement using modified IMCRA and OMLSA methods," in *International Conference on Communications and Electronics 2010*, pp. 195–200, Nha Trang, Vietnam, 2010.

[17] L. Xu, Z. Wei, S. F. A. Zaidi, B. Ren, and J. Yang, “Speech enhancement based on nonnegative matrix factorization in constant-Q frequency domain,” *Applied Acoustics*, vol. 174, article 107732, 2021.

[18] S. Thakallapalli, S. Gangashetty, and N. Madhu, “A new weighted NMF algorithm for missing data interpolation and its application to speech enhancement,” in *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, A Coruna, Spain, 2019.

[19] Y. Xiang, L. M. Shi, J. L. Hojvang, M. H. Rasmussen, and M. G. Christensen, “A noval NMF-HMM speech enhancement algorithm based on Poisson mixture model,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, Toronto, ON, Canada, 2021.

[20] S. Lee, D. K. Han, and H. Ko, “Single-channel speech enhancement method using reconstructive NMF with spectrotemporal speech presence probabilities,” *Applied Acoustics*, vol. 117, pp. 257–262, 2017.

[21] H. R. Jia, W. M. Wang, and S. L. Mei, “Combining adaptive sparse NMF feature extraction and soft mask to optimize DNN for speech enhancement,” *Applied Acoustics*, vol. 171, article 107666, 2021.

[22] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, “NMF-based target source separation using deep neural network,” *IEEE Signal Process Letters.*, vol. 22, no. 2, pp. 229–233, 2015.

[23] S. Nie, S. Liang, W. Liu, X. Zhang, and J. Tao, “Deep learning based speech separation via NMF-style reconstructions,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2043–2055, 2018.

[24] M. Sun, Y. N. Li, J. F. Gemmeke, and X. Zhang, “Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 33, no. 7, pp. 1233–1242, 2015.

[25] Y. H. Lai, S. S. Wang, C. H. Chen, and S. H. Jhang, “Adaptive Wiener gain to improve sound quality on nonnegative matrix factorization-based noise reduction system,” *IEEE Access*, vol. 7, pp. 43286–43297, 2019.

[26] S. Zhang, J. Tian, A. Banerjee, and J. Li, “Automatic recognition of porcine abnormalities based on a sound detection and recognition system,” *Transactions of the ASABE*, vol. 62, no. 6, pp. 1755–1765, 2020.

[27] C. D. Xu, J. Zhou, D. G. Ying, and P. Xin, “Heart sound denoising by dynamic noise estimation,” *Journal of Biomedical Engineering*, vol. 37, no. 5, pp. 775–785, 2020.

[28] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[29] J. Wang, S. Guan, S. Liu, and X. L. Zhang, “Minimum-volume multichannel nonnegative matrix factorization for blind audio source separation,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 3089–3103, 2021.

[30] A. Parimala Gandhi and S. Vijayan, “Upgrading sparse NMF algorithm for blind source separation through adaptive parameterized hybrid kernel based approach,” *Measurement*, vol. 143, pp. 11–21, 2019.

[31] D. D. Lee and H. S. Seung, “NMF-based target source separation using deep neural network,” *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, 2015.

[32] J. Erkelens, J. Jensen, and R. Heusdens, “A data-driven approach to optimizing spectral speech enhancement methods for various error criteria,” *Speech Communication*, vol. 49, no. 7-8, pp. 530–541, 2007.

[33] J. F. Gemmeke, D. P. Ellis, D. Freedman et al., “Audio set: an ontology and human-labeled dataset for audio events,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780, New Orleans, LA, USA, 2017.

[34] W. E. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, “The DARPA speech recognition research database: specifications and status,” *Proc DARPA Speech Recognition Workshop*, pp. 93–99, 1986.

[35] V. Andrew and J. M. S. Herman, “Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[36] K. J. Piczak, “ESC: dataset for environmental sound classification,” in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, Brisbane, Australia, 2015.

[37] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” *2001 IEEE international conference on acoustics, speech, and signal processing*, vol. 2, pp. 749–752, 2001.

[38] I. M. Hameed, S. H. Abdulhussain, and B. M. Mahmmod, “Content-based image retrieval: a review of recent trends,” *Cogent Engineering*, vol. 8, no. 1, 2021.

[39] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences*, vol. 6, no. 6, pp. 162–179, 2016.

[40] L. Y. Luo, L. J. Zhang, M. Wang et al., “A system for the detection of polyphonic sound on a university campus based on CapsNet-RNN,” *IEEE Access*, vol. 9, pp. 147900–147913, 2021.

[41] M. Wang, M. Yao, L. Luo et al., “Environmental sound recognition based on double-input convolutional neural network model,” in *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, pp. 620–624, Weihai, China, 2020.