ISSN 1064-2269, Journal of Communications Technology and Electronics, 2016, Vol. 61, No. 6, pp. 672–678. © Pleiades Publishing, Inc., 2016.
Original Russian Text © Y. Sandoval-Ibarra, V.H. Diaz-Ramirez, V.I. Kober, V.N. Karnaukhov, 2015, published in Informatsionnye Protsessy, 2015, Vol. 15, No. 3, pp. 314–323.

Speech Enhancement with Adaptive Spectral Estimators

Y. Sandoval-Ibarra[a], V. H. Diaz-Ramirez[a], V. I. Kober[b, c], and V. N. Karnaukhov[c]

[a] *Instituto Politécnico Nacional-CITEDI, Ave. del Parque 1310, Mesa de Otay, Tijuana B.C. 22510, México*
[b] *Departament of Computer Science, CICESE, Carretera Ensenada-Tijuana 3918,*
Zona Playitas, Ensenada, B.C. 22860, México
[c] *Institute for Information Transmission Problems, Russian Academy of Sciences,*
Bol'shoi Karetnyi per. 19 str. 1, Moscow, 127051 Russia
e-mails: vnk@iitp.ru, vkober@cicese.mx

Received June 8, 2015

Abstract—Common statistical estimators for speech enhancement rely on several assumptions about statistical properties of speech and noise processes. In real applications, these assumptions may not be always satisfied due to the effects of a nonstationary environment. In this work, we propose new robust spectral estimators for speech enhancement by incorporation of calculation of rank-order statistics to existing speech enhancement estimators. The proposed estimators are better adapted to nonstationary characteristics of speech signals and noise processes in real environments. By means of computer simulations, we show that the proposed estimators outperform the known estimators in terms of objective criteria of quality.

Keywords: speech filtering, speech enhancement, robust estimators

DOI: 10.1134/S1064226916060218

INTRODUCTION

Modern methods of speech enhancement rely on the optimization of metrics (such the quality and intelligibility [1]) well describing the subjective perception of reconstructed speech signals. Telecommunication systems are examples of applications whose work requires reliable algorithms providing high intelligibility of speech with reduced noise. The speech enhancement is usually formulated as the problem of estimating the magnitude spectrum of clean speech signal from the observed signal. There are several successful estimators for speech enhancement, among which it is worth mentioning the following ones: maximum likelihood (ML) [1–3], minimum mean-squared error (MMSE) [4–6], logarithmic minimum mean-squared error (log-MMSE) [7], and maximum a posteriori probability (MAP) [8]. These estimators rely on several assumptions about the statistical properties of speech signals and noise. For example, it is common practice to assume an asymptotic behavior of the statistical characteristics of speech signals and a known distribution of the speech signal. In real applications, the local distribution density of speech signals and noise can change with time. Therefore, the existing estimators can lead to unsatisfactory results when processing real speech signals on a nonuniform environmental background. Thus, the design of locally adaptive robust estimators for speech enhancement is desirable.

Digital signal processing widely employs filters based on the calculation of rank-order statistics [9–11]. Such filters are robust to the heavy tailed noise and preserve fine details and rapid changes in the signal. These properties of nonlinear filters are useful in the speech enhancement for the suppression of undesirable noise while preserving the intelligibility of speech. Recently, a locally adaptive nonlinear filtering for speech processing was proposed [12], which is capable of reducing the additive noise and preserving the intelligibility of speech almost without artifacts such as a "musical" noise. However, the obtained estimator does not take into account the metrics describing the subjective perception of speech signals by a man [13, 14]. In the present work, we propose the use of rank-order statistics for improvement of the existing estimators of speech processing taking into account the subjective perception of speech signals.

Let a discrete function $f(n) = s(n) + d(n)$ be an input speech fragment of length N, the function $s(n)$ is the undistorted signal, and $d(n)$ is an additive noise with the zero mean. In the frequency domain, the observed signal can be represented as

$$F_k e^{j\theta_k^f} = S_k e^{j\theta_k^s} + D_k e^{j\theta_k^d}, \tag{1}$$

where F_k, S_k, and D_k are the magnitudes of the discrete Fourier spectra of the signals $f(n)$, $s(n)$, and $d(n)$, respectively, and θ_k^f, θ_k^s, and θ_k^d are the phases of the observed, undistorted signals and the noise, respectively. After obtaining an estimate of the spectrum of the undistorted signal from the observed signal, the reconstructed spectrum of the signal is calculated as $S_k e^{j\theta_k^s} \approx \hat{S}_k e^{j\theta_k^f}$.

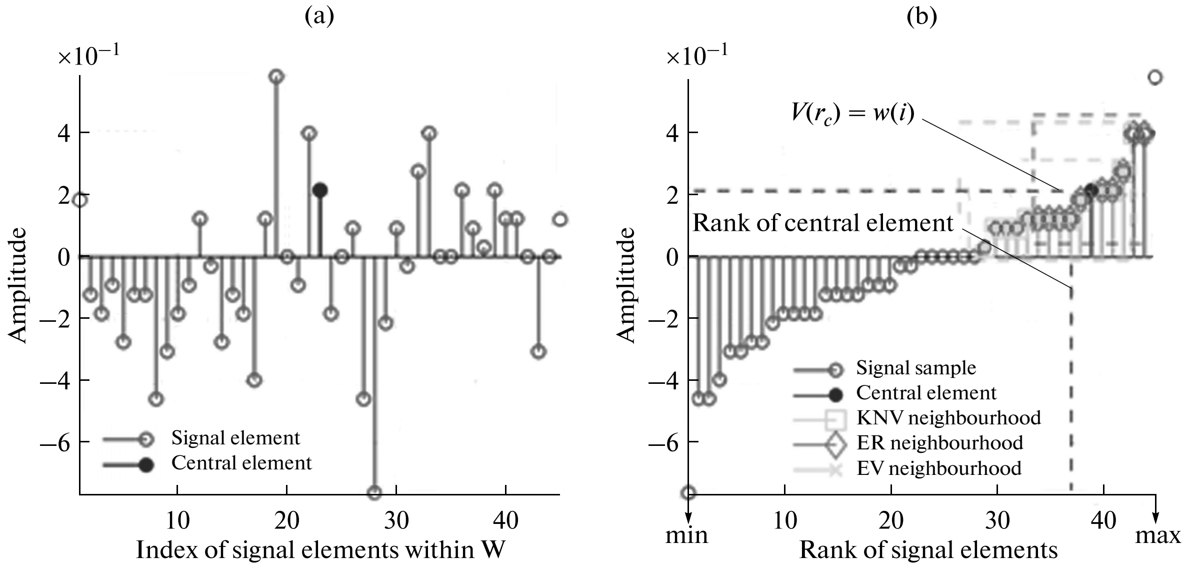

Fig. 1. Calculation of locally adaptive neighborhoods: (a) sliding window and (b) variation row and local neighborhoods.

The existing estimators of speech spectrum presume that the Fourier coefficients (the real and imaginary parts) of speech signals and noise have zero means, and the signals are independent Gaussian random processes quasi-stationary in the interval of 20–40 milliseconds. In real applications, these assumptions may be wrong, e.g., when the speech signal is distorted by a non-Gaussian or nonstationary noise. In this work, we propose a robust estimator based on calculation of rank-order statistics for speech enhancement. The proposed estimates can be adapted to nonstationary features of noised speech signals. They can also enhance speech without loss of intelligibility and without introducing artifacts to the signal.

The structure of the paper is as follows. Section 1 outlines the locally adaptive speech processing with the use of rank-order statistics. Section 2 describes the proposed algorithm for speech enhancement. Section 3 presents experimental results obtained with the help of the proposed approach. These results are compared in objective criteria with the results obtained by the known methods. The final section presents our conclusions.

1. SPEECH ENHANCEMENT WITH LOCAL SIGNAL PROCESSING

The design of rank-order filters is usually performed in two stages: at first, local uniform neighborhoods are singled out (the structural approach) and, then, estimates of the undistorted signal (estimators) are constructed [10, 11]. The locally adaptive signal processing is performed in a sliding window. In the first step, homogeneous neighborhoods in a sliding window—the desired structures of the signal in a window—are defined. Then, on the basis of the defined elements of the local neighborhoods, an estimate of the undistorted signal for the central element of the window with the chosen criterion is constructed. Figure 1 shows an example of a speech signal and construction of local neighborhoods based on rank-order statistics.

For the processed speech signal, the vector of a sliding window $\mathbf{w}$ consisting of S elements can be represented as follows:

$$\mathbf{w} = \left[w\left(n - i + \frac{S+1}{2}\right) = f(n) : |n - i| \le \frac{S-1}{2}\right]^T, \tag{2}$$

where i is the index of the central element within the current window and T denotes transposition. The variation row $v(r)$ is an ordered sequence of elements of the vector $\mathbf{w}$ satisfying the following condition: $v(1) \le v(2) \le \ldots \le v(S)$. The quantities $v(r)$ and $r(v)$ are the rth rank-order statistics and the rank of the quantity v, respectively [10]. It should be noted that the rank-order statistics and the rank can be calculated from the local histogram of the signal $\{h(q), q = 0, \ldots, Q - 1\}$ inside the sliding window as $r(v) = \sum_{q=0}^{v} h(q)$, where Q is the number of signal quantization levels.

There are several variants of constructing local neighborhoods based on rank-order statistics [10]. One of the most popular neighborhoods for signal processing is the EV-neighborhood. This neighborhood is the subset of elements of the vector $\mathbf{w}$ whose values deviate from the central element $w(i)$ at least by the specified values $-\varepsilon_v$ and $+\varepsilon_v$ as follows:

$$\mathbf{v} = \{v(n) = w(n) : w(i) - \varepsilon_v \le w(n) \le w(i) + \varepsilon_v\}, \tag{3}$$

where $\mathbf{v}$ is the vector $S_a \times 1$ $(S_a \le S)$ whose elements form the subset of element of the vector $\mathbf{w}$. Estimators for speech enhancement are constructed using popular methods of statistical estimation [3]. For example,

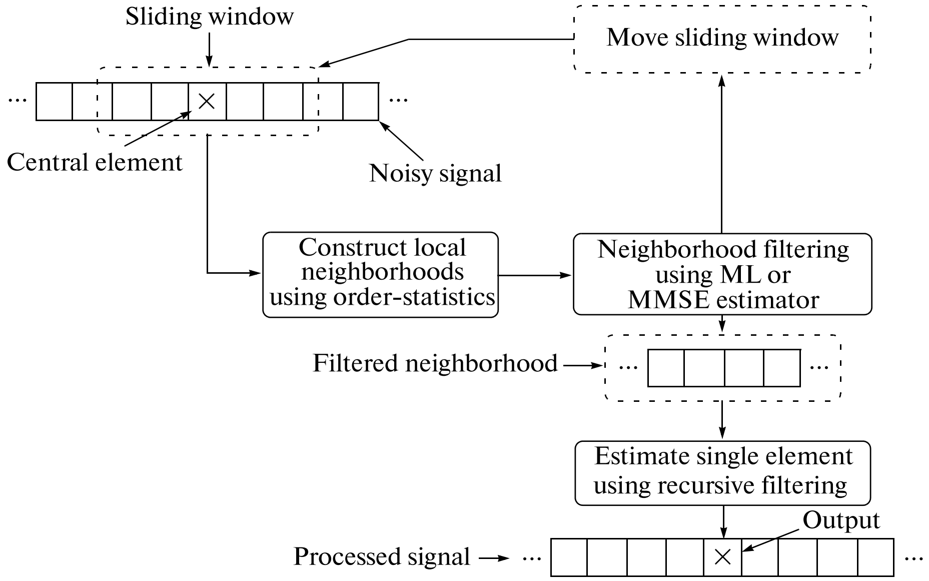
Sliding window
Move sliding window
Central element
Noisy signal
Construct local neighborhoods using order-statistics
Neighborhood filtering using ML or MMSE estimator
Filtered neighborhood
Estimate single element using recursive filtering
Output
Processed signal

Fig. 2. Block diagram of the proposed rank filtering.

in [4], optimal estimates for the spectrum magnitude of undistorted speech signal with respect to the mean-squared error and logarithmic mean-squared error were proposed. These estimators usually give good results for the suppression of a stationary noise in a speech signal. On the other hand, the use of these estimators deteriorates the subjective quality of speech signal, because the reconstructed signal contains an annoying "musical" noise, which can harass the listener.

The MMSE estimator of the spectrum magnitude of undistorted speech signal [4] can be written as

$$\hat{S}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\beta_k}}{\gamma_k}\exp\left(-\frac{\beta_k}{2}\right) \times\left[(1+\beta_k)I_0\left(\frac{\beta_k}{2}\right)+\beta_k I_1\left(\frac{\beta_k}{2}\right)\right]V_k, \tag{4}$$

where I_0 and I_1 are the modified Bessel functions of the zero and first orders, respectively, and β_k is calculated as

$$\beta_k = \frac{\xi_k}{1+\xi_k}, \tag{5}$$

where ξ_k and λ_k are a priori and a posteriori signal-to-noise ratios (SNR), given by

$$\gamma_k = \frac{V_k^2}{1+\lambda_k^d} \tag{6}$$

and

$$\xi_k = \frac{\lambda_k^s}{\lambda_k^d}. \tag{7}$$

Here, λ_k^s and λ_k^d are the variances of the undistorted signal and noise, respectively. In the time domain, the central element of the sliding window of the undistorted signal can be calculated as follows:

$$y(i) - ay_{k-1}(i) + (1-a)y_k(i), \tag{8}$$

where $a \in [0, 1]$ is a weight coefficient and $y_k(i)$ is obtained from

$$y_k(i) = \mu_s + \frac{\lambda_k^s}{\lambda_k^s+\lambda_k^d}(s(i)-\mu_s), \tag{9}$$

where μ_s is the mean of the reconstructed signal and $s(i)$ is the central element of the sliding window after applying the estimate of the minimum mean-squared error.

In this work, we propose the method for processing of speech signals by modifying the existing estimators to enhance speech while preserving the intelligibility and without introducing artificial sounds.

2. THE PROPOSED ALGORITHM

In this section, we describe the proposed algorithm for speech enhancement by locally adaptive signal processing. The block diagram of the algorithm is presented in Fig. 2, and its steps are detailed below.

Step 1. Read the initial input segment $\mathbf{n}_0$ with S elements in the absence of a speech signal.

Step 2. Read the initial input speech segment $f(n)$ with N elements and set $i = 1$.

Step 3. Create the vector of window $\mathbf{w}$ around the ith noised element, using expression (2).

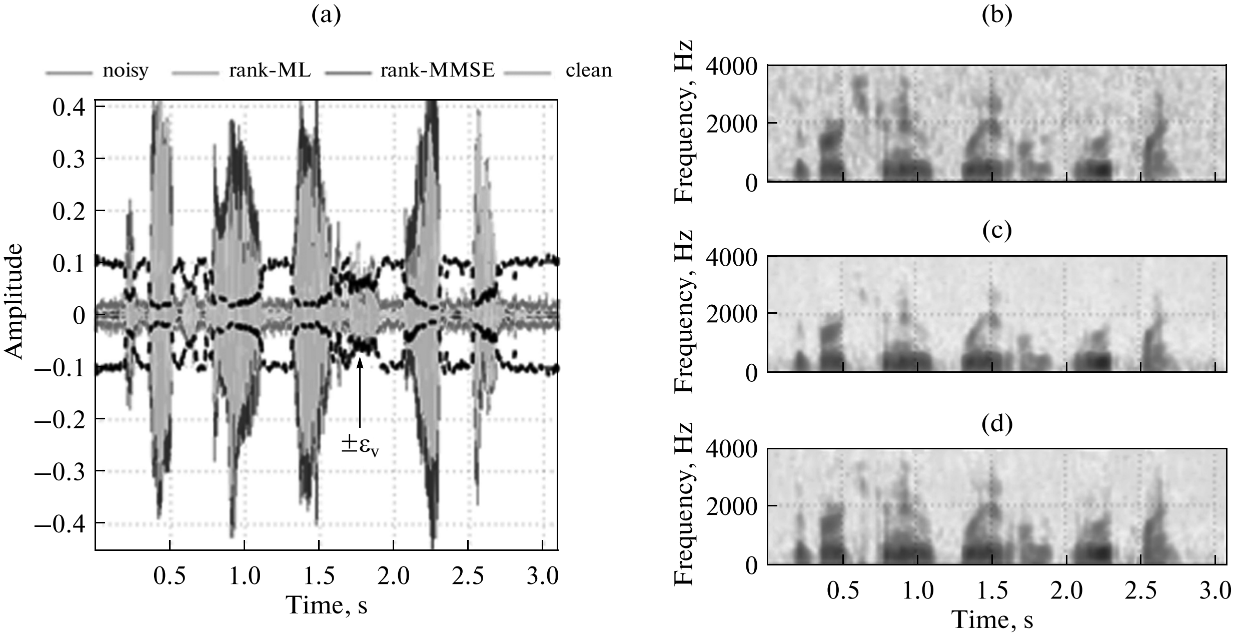

Fig. 3. Example of enhancement of speech distorted by an additive Gaussian noise with SNR = 15 dB by the proposed algorithms: (a) speech signals, (b) spectrogram of a noised signal, (c) spectrogram of the signal processes with the rank-ML algorithm, and (d) spectrogram of the signal processes with the rank-MMSE algorithm.

Step 4. Calculate the value of ε_v as follows [12]:

$$\varepsilon_v = \alpha_1\sigma_f\left[1-\frac{1}{1-\Omega^{-\alpha_2}}\right], \qquad (10)$$

where Ω is the local SNR calculated by the formula $\Omega = \dfrac{\mathbf{w}^T\mathbf{w}}{\mathbf{n}_0^T\mathbf{n}_0}$, and σ_f is the standard deviation of the noise. The parameters $a_1 \geq 1$ and $a_2 \in (0, 1]$ take into account a priori information on the spread of the speech signal and fluctuation of noise.

Step 5. Construct *EV*-neighborhood **v** of the vector **w** by expressions (3) and (10).

Step 6. Apply the estimate of the minimum mean-squared error, using expression (4).

Step 7. Calculate the estimate of the input signal by expressions (8) and (9). Set $i = i + 1$. If $i \leq N_i$, then go to Step 3, else go to Step 2.

The result of work of the algorithm is the signal reconstructed with the optimal mean-squared estimate of the undistorted signal and the locally adaptive processing based on rank-order statistics.

3. EXPERIMENTAL RESULTS

In this section, we present experimental results obtained by the proposed method. Numerous experiments were performed for testing the quality of the method. The results were compared with those obtained by existing algorithms for speech enhancement. We tested the classical maximum-likelihood algorithm, denoted by ML [1–3], the minimum mean-squared error algorithm, denoted by MMSE [4], and the proposed algorithms based on the rank-order statistics and estimators of the mean-squared error and maximum likelihood, denoted as rank-MMSE and rank-ML, respectively.

All the algorithms were tested on speech signals distorted by two types of noise: the white normal noise and automobile noise. The SNR tooks the values of 20, 15, and 10 dB. The considered algorithms were tested using the IEEE data base [13]. This base contains 600 speech sentences pronounced by male and female speakers. The sentences in the data base are phonetically balanced and have a relatively low predictability of the vocabulary context. The sentences were recorded with the sampling rate of 8 kHz. The quality of work of the algorithms was estimated using the following metrics.

—The quality of speech is characterized by the Perceptual Evaluation of Speech Quality (PESQ) [14];

—The intelligibility of speech is estimated by the short-time objective intelligibility (STOI) [15];

—The noise suppression is characterized by the useful source-to-interference ratio (SIR) [16];

—The addition of artifacts to the speech signal is described by the useful source-to-artifact ratio (SAR) [16].

Let us test the quality of the algorithms on a speech signal distorted by an additive noise. The parameters for the proposed algorithms are as follows: $S = 121$, $a_1 = 2.0$, $a_2 = 0.45$, and $a = 0.8$. Figure 3 shows an example of speech enhancement by the proposed

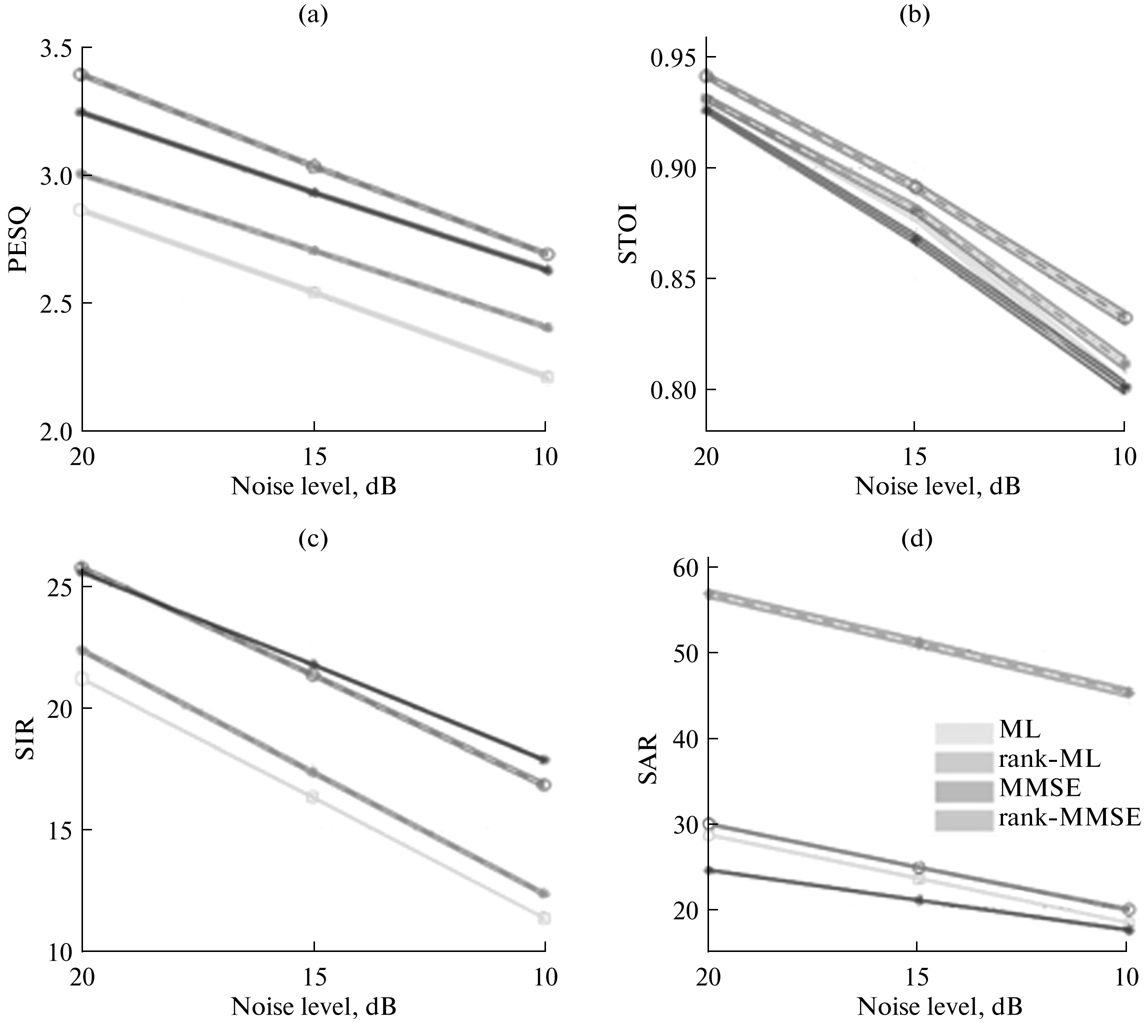

Fig. 4. Results of processing of speech distorted by an additive Gaussian noise with SNR = 15 dB by the proposed algorithms with 95% confidence in terms of (a) the perceptual evaluation of speech quality (PESQ), (b) short-time objective intelligibility (STOI), (c) noise reduction (SIR), and (d) introduction of artifacts (SAR).

algorithms when the speech is distorted by an additive noise with the SNR = 15 dB. In addition, Fig. 3 shows spectrograms of the noised and processed signals. Note how ε_v is adapted to the local variations in the noised signal: when the local SNR is low, ε_v takes large values. This means that, in this case, the proposed algorithms perform more aggressive filtration than at large SNR. Figure 4 presents the results of processing of 600 distorted speech signals by the tested algorithms with a 95% confidence in terms of the quality of speech, intelligibility, noise suppression, and introduction of artifacts. It should be noted that the proposed rank-ML algorithm gives better results in the quality of speech and noise suppression than the classical ML algorithm for all considered values of the SNR. We can also see that the proposed rank-ML algorithm outperforms all the tested algorithms with respect to the SAR. The proposed rank-ML algorithm outperforms the common algorithms in terms of the perceptual quality of speech and the source-to-artifact ratio for all considered values of the SNR. Moreover, the rank-MMSE algorithm is the best one of all tested algorithms in the intelligibility for all values of the SNR. The classical MMSE algorithm suppresses noise well due to introducing a substantial "musical" noise (the worst values of the SAR).

Now let us analyze the quality of work of the speech enhancement algorithms in an environment with automobile noise. The parameters of the proposed algorithms for this experiment were as follows: $S = 65$, $a_1 = 2.0$, $a_2 = 0.5$, and $a = 0.8$. Figure 5 shows and example of the enhancement of speech distorted by an automobile noise with the SNR = 15 dB and the spectrograms of the noisy and processed signals. Figure 6 presents the results of processing of 600 distorted speech signals by the tested algorithms with a 95% confidence in terms of the quality of speech, intelligibility, noise suppression, and introduction of artifacts.

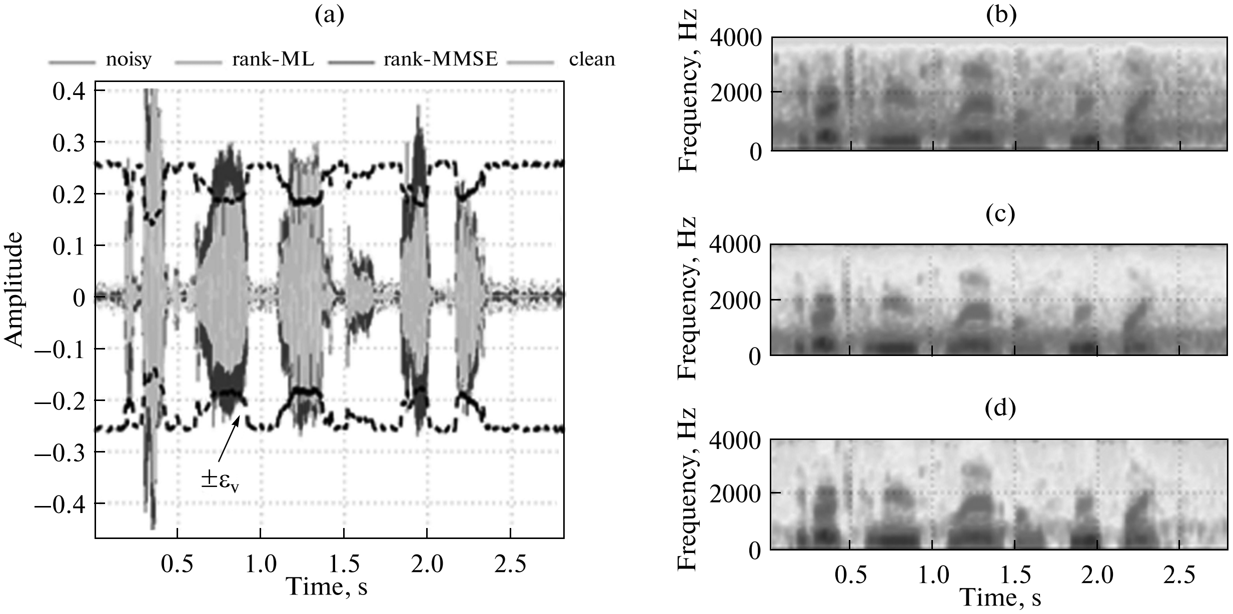

Fig. 5. Example of enhancement of speech distorted by an automobile noise with SNR = 15 dB by the proposed algorithms: (a) speech signals, (b) spectrogram of noised signal, (c) spectrogram of the signal processes with the rank-ML algorithm, and (d) spectrogram of the signal processes with the rank-MMSE algorithm.

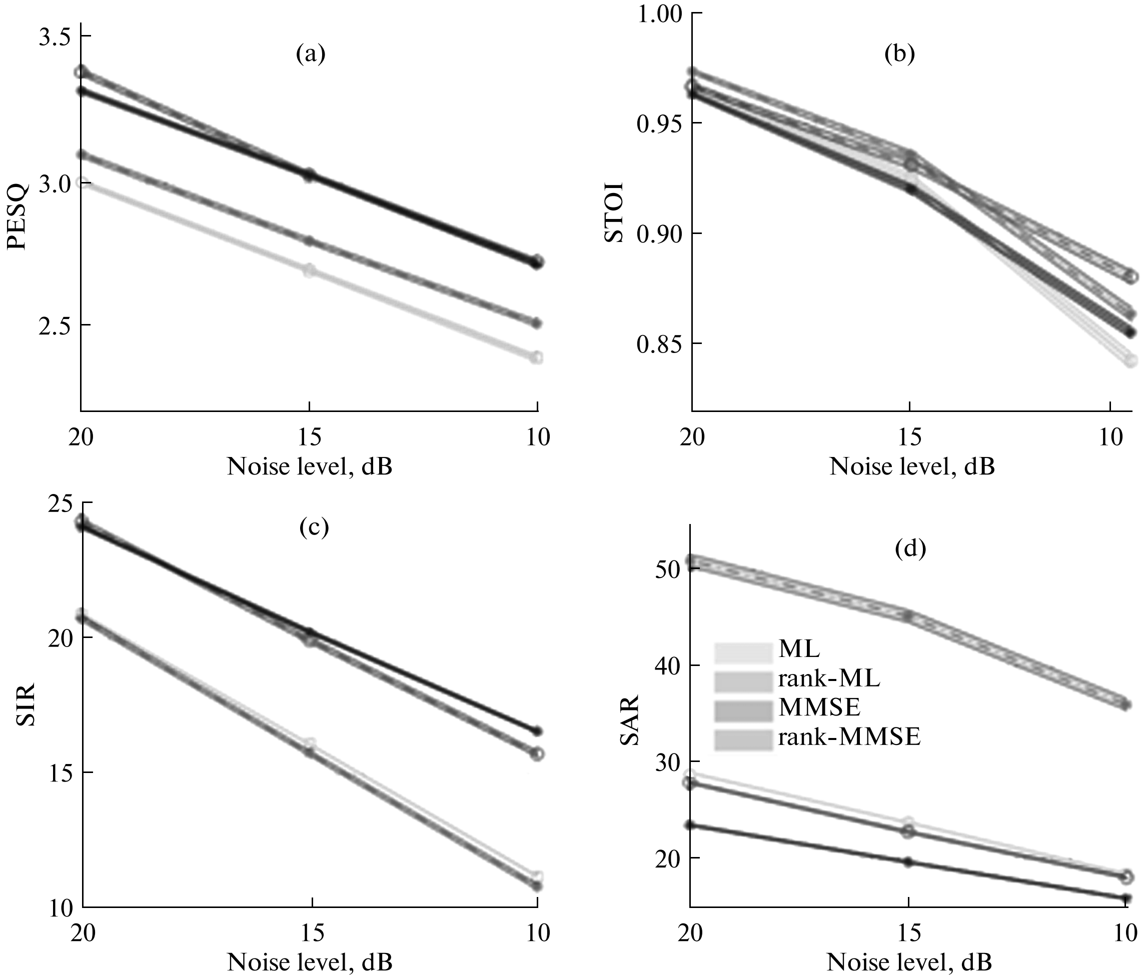

Fig. 6. Results of processing of speech distorted by an automobile noise with SNR = 15 dB by the tested algorithms with 95% confidence in terms of (a) the perceptual evaluation of speech quality (PESQ), (b) short-time objective intelligibility (STOI), (c) noise reduction (SIR), and (d) introduction of artifacts (SAR).

It should be noted that the proposed rank-ML algorithm provides a substantial enhancement of speech with respect to the perceptual quality of speech and noise suppression than the classical ML algorithm for all considered values of the SNR. It should also be noted that the noise suppression of the rank-ML algorithm is similar to that of the classical MMSE algorithm. The common ML algorithm is the worst with respect to the perceptual quality of speech and noise reduction among all tested algorithms. The proposed rank-ML algorithm outperforms all the algorithms in the perceptual quality of speech and noise suppression for the SNR of 15 and 20 dB. The common MMSE algorithm yields slightly better performance and noise suppression when the speech signal is strongly noised to the SNR = 10 dB. However, as in the previous results with an additive noise, the MMSE algorithm has the worst result among all tested algorithm in terms of the SAR criterion.

CONCLUSIONS

In this paper, we have proposed new algorithms for speech enhancement based on known estimates of spectrum amplitudes of distorted signals and rank-order statistics. The proposed estimators are easily adapted to nonstationary characteristics of speech signals and background noise in real conditions. By computer simulation, it was shown that the proposed algorithms for speech enhancement outperform the classical methods in terms of objective criteria of quality.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education and Science of Russian Federation (grant 2.1766.2014K).

REFERENCES

1. P. Loizou, *Speech Enhancement: Theory and Practice*, 2nd Ed. (Taylor & Francis, Boca Raton, 2013).
2. R. McAulay and M. Malpass, “Speech enhancement using a softdecision noise suppression filter,” IEEE Trans. Acoust., Speech, Signal Process. **28**, 137–145 (1980).
3. N. S. Kim and J.-H. Chang, “Spectral enhancement based on global soft decision,” IEEE Signal Process. Lett. **7** (5), 108–110 (2000).
4. Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process. **32**, 1109–1121 (1984).
5. G.-H. Ding, T. Huang, and B. Xu, “Suppression of additive noise using a power spectral density MMSE estimator,” IEEE Signal Process. Lett. **11**, 585–588 (2005).
6. R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” IEEE Trans. Speech Audio Process. **13**, 845–856 (2005).
7. Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process. **33**, 443–445 (1985).
8. T. Lotter and P. Vary, “Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model,” EURASIP J. Appl. Signal Process., No. 7, 1110–1126 (2005).
9. L. Yaroslavsky and M. Eden, *Fundamentals of Digital Optics* (Birkhaeuser, Boston, 1996).
10. V. Kober, M. Mozerov, and J. Alvarez-Borrego, “Nonlinear filters with spatially connected neighborhoods,” Opt. Eng. **40**, 971–983 (2001).
11. P. J. Huber, P. C. Pop, and E. M. Ronchetti, *Robust Statistics*, 2nd Ed., (Wiley, New York, 2009).
12. V. M. Diaz-Ramirez and V. Kober, “Robust speech processing using local adaptive nonlinear filtering,” IET Signal Process. **7**, 345–359 (2013).
13. “IEEE Subcommittee (1969), IEEE recommended practice for speech quality measurements,” IEEE Trans. Audio Electroacoust. **17**, 225–246 (1969).
14. “ITU, Perceptual evaluation of speech quality (PESQ). An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” ITU-T Recommendation, 862 (2001).
15. C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. Audio, Speech, Language Process. **19**, 2125–2136 (2011).
16. E. Vincent, R. Gribonval, and C. F‘evotte, “Performance measurement in blind audio source separation,” IEEE Trans. Audio, Speech, Language Process. **14**, 1462–1469 (2006).

Translated by E. Chernokozhin