Speech Enhancement with Local Adaptive Rank-Order Filtering

Vitaly Kober[1], Victor Diaz Ramirez[2], and Yuma Sandoval Ibarra[2]

[1] Computer Science Department, CICESE, Ensenada, B.C., Mexico

[2] Instituto Politécnico Nacional, CITEDI, Tijuana, B.C., Mexico

vkober@cicese.mx, vdiazr@ipn.mx, juma_san@hotmail.com

Abstract. A local adaptive algorithm for speech enhancement is presented. The algorithm is based on calculation of the rank-order statistics of an input speech signal over a moving window. The algorithm varies the size and contents of a sliding window signal as well as an estimation function employed for recovering a clean speech signal from a noisy signal. The algorithm improves the quality of a speech signal preserving its intelligibility. The performance of the algorithm for suppressing additive noise in an input test speech signal is compared with that of common speech enhancement algorithms in terms of objective metrics.

Keywords. Speech enhancement, local adaptive filtering, rank-order statistics, musical noise, intelligibility.

Mejora de voz con filtrado local adaptativo basado en estadísticas de orden

Resumen. Se presenta un algoritmo localmente adaptativo para la mejora de voz. El algoritmo, se basa en el cálculo de estadísticas de orden prioritario de una señal de voz dentro de una ventana deslizante. El algoritmo es localmente adaptativo ya que puede variar el tamaño y contenido de la señal dentro de la ventana deslizante así como también, la función de estimación usada para la recuperación de la señal limpia a partir de la señal ruidosa. El algoritmo propuesto mejora la calidad de la voz preservando la inteligibilidad del mensaje, e introduciendo únicamente ruido musical imperceptible. El desempeño del algoritmo propuesto es comparado con el desempeño de los algoritmos existentes en términos de varias métricas objetivas.

Palabras clave. Mejora de voz, filtrado local adaptativo, estadísticas de orden prioritario, ruido musical, inteligibilidad.

1 Introduction

Speech enhancement has received research interests owing to recent developments in modern communication equipment such as smart phones, voice over IP and teleconference systems, and speech recognition devices among others [1, 33]. The task of speech enhancement consists in improving the quality of a captured voice signal in terms of noise reduction with respect to different performance criteria [2, 3, 4]. Speech enhancement is a difficult task because voice signals are highly time-variant. Moreover, signals corrupted by ambient noise can be described by a mixture of several random processes with different statistical distributions [5]. Furthermore, common speech enhancement algorithms introduce a typical distortion to processed signals that considerably reduces speech intelligibility [2]. For this reason, an optimization of the noise-suppression to signal-distortion ratio is desirable [3]. Actually, speech enhancement can be broadly classified in two groups: single-channel and multiple-channel based systems. Single-channel systems utilize only one microphone to capture speech signals, whereas multiple-channel systems use a microphone array to better characterize and suppress noise [5, 6]. In this paper we focus on single-channel systems. Two common techniques adopted in single-channel speech enhancement are the Wiener filtering and the spectral subtraction based methods; these algorithms are carried out in the frequency domain. The Wiener filtering [7, 8] is a linear filter optimized with respect to the mean-squared-error between the clean and processed signals. This

ISSN 1405-5546
http://dx.doi.org/10.13053/CyS-18-1-2014-023

formulation requires an estimate of the clean speech power spectrum, which can be obtained iteratively [9, 10] or non-iteratively [8, 7] from a noisy speech. On the other hand, spectral subtraction methods require only an estimate of the average noise spectrum which is subtracted from the spectrum of the captured signal. In such a manner, the average signal-to-noise ratio (SNR) is improved [11, 12]. Since both strategies modify the spectral distribution of speech signals, they commonly introduce artificial artifacts to processed signals such as musical noise [6]. Various proposals were suggested to alleviate this drawback [13].

It is important to note that majority of existing speech enhancement algorithms assume that noise functions corrupting speech signals are stationary [7]. However, this assumption is not true for real applications, for instance, when speech is corrupted by a mixture of background noise and impulsive-like noise caused by imperfections in communication channels [14]. In this case the use of a robust filtering is desirable [15]. Usually, impulsive noise in speech signals is suppressed with a two-step procedure, that is, impulsive outliers are first detected and removed followed by an estimation of speech samples using interpolation [5]. Note that this approach is not effective when speech is corrupted by a mixture of additive and impulsive noise. In signal processing there exist several successful nonlinear filters based on calculation of rank-order statistics [16, 17, 18]. These filters are robust and able to preserve fine details of signals comparing with those of conventional linear filtering. In speech enhancement, this feature can help to suppress the background noise introducing only imperceptible musical noise and, therefore, to preserve intelligibility of processed speech. The median filtering [19], multilevel and multistage median filters [20], stack filtering [18], alpha-trimmed mean filters [17], and rank-order filters [21] are important nonlinear filters that have been proved to be very effective in suppression of additive and impulsive noise. We are interested in designing a robust rank-order filter for local adaptive processing of a noisy speech signal. Basically, taking into account signal and noise models, a robust nonlinear filter can be designed by optimizing some performance criteria [22, 15]. In this work, we propose a locally adaptive algorithm for robust speech enhancement which is carried out in a sliding window. The algorithm is able to vary the size and contents of the sliding window as well as an estimation function employed for recovering a clean speech signal from a signal to be processed. This means that a noise-free signal can be recovered with a time-variant estimator over a moving adaptive-window. The proposed algorithm is adaptive to nonstationary signal fragments and noise fluctuations. As a result, the method improves the speech quality with a low level of musical noise. Note that since the proposed method is implemented in the time domain, then, a priori estimation of the SNR using modified spectral distributions of signals and noise is not carried out for each frame. It is known that the decision-directed based methods use such estimation that introduces annoying artifacts in processed speech [23].

The paper is organized as follows. Section 2 provides a review of rank-order processing and presents the proposed adaptive algorithm for speech enhancement. Section 3 illustrates the results obtained with the proposed algorithm. The performance of the algorithm is compared with that of existing methods for speech enhancement in terms of several objective performance metrics. Finally, Section 4 summarizes our conclusions.

2 Speech Enhancement with Local Adaptive Rank-Order Filtering

Rank-order filtering is a locally adaptive signal processing carried out in a sliding window. First, local neighborhoods are used to define desirable data-structures in the window. Then, a robust estimation is applied to elements of the neighborhood for computing only an estimate of the central element of the window. A block diagram of basic rank-order speech processing is presented in Fig. 1. A local neighborhood is a subset of signal elements of the sliding window which are close in some sense to a given element [16]. Note that there exist several options to construct local neighborhoods; however, for robust estimation the use of order-statistics is a good choice [16, 22]. Order-statistics can exploit a relationship between window elements previously sorted in the ascending order with respect to their values.

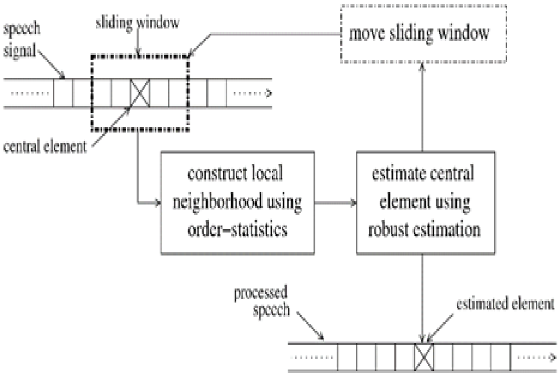

speech signal
sliding window
move sliding window
central element
construct local neighborhood using order–statistics
estimate central element using robust estimation
processed speech
estimated element

Fig. 1. Block diagram of basic rank-order filtering

Let $f_j(n) = s_j(n) + \left[1 - p_j(n)\right] b_j(n) + p_j(n) d_j(n)$ be the *j*th input speech segment to be processed, with $n = 1,\ldots,N,$ the signal range of [-1,1], and *Q* quantization levels. Note that real speech signals have infinite duration, $j = 1,\ldots,\infty$. In the signal model, $s_j(n)$ is an uncorrupted speech segment to be estimated, $b_j(n)$ is a zero-mean additive noise, $d_j(n)$ represents impulsive noise, and $p_j(n)$ is a binary random function defined as one when the impulsive noise is present and zero otherwise. The input speech segment can be rewritten in a matrix-vector notation as follows:

$$\mathbf{f}_j = \mathbf{s}_j + \left(\mathbf{I} - \mathbf{P}_j\right)\mathbf{b}_j + \mathbf{P}_j \mathbf{d}_j, \tag{1}$$

where $\mathbf{f}_j$, $\mathbf{s}_j$, $\mathbf{b}_j$, and $\mathbf{d}_j$ are $N \times 1$ vectors which represent the discrete sequences $f_j(n)$, $s_j(n)$, $b_j(n)$, and $d_j(n)$, respectively. In addition, $\mathbf{I}$ is the $N \times N$ identity matrix and $\mathbf{P}_j$ is a $N \times N$ diagonal matrix whose entries are the elements of the sequence $p_j(n)$. For *i*th position of the running window, a sliding window vector $\mathbf{w}_{j,i}$ with *S* elements can be constructed as

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

where *S* is the size of the vector (odd number), *i* is the index of the central window element, and *T* denotes transpose. In order to avoid boundary effects the input speech segment may be overlapped. Local neighborhoods are constructed from the sliding window elements by sorting them in the ascending order with respect to their values. The ordered sequence is a variational row, and it is defined as $\left\{ V(r); r = 1,\ldots,S \right\}$, where $V(1) \le V(2) \le \cdots \le V(S)$. The quantities $V(r)$ and $r(V)$ are called the *r*th order-statistics and the rank of the value *V*, respectively [16, 21]. These quantities can be computed from the histogram of the sliding window data $\left\{ h(q); q = 0,\ldots,Q-1 \right\}$, as $r(V) = \sum_{q=0}^{V} h(q)$. Let $w_{j,i}(i)$ be the central element of the sliding window placed at the *i*th position. We introduce the EV-neighborhood which describe convenient relationships between signal elements [16, 21]. The EV-neighborhood is a subset of $\mathbf{w}_{j,i}$ elements whose values deviate from the value of the central element $w_{j,i}(i)$ at most by prespecified quantities $-\varepsilon_V$ and $+\varepsilon_V$. This neighborhood can be constructed from the sliding window as

$$\mathbf{v}_{j,i} = \Big[v_{j,i}(n) = w_{j,i}(n) : \\ w_{j,i}(i) - \varepsilon_V \le w_{j,i}(n) \le w_{j,i}(i) + \varepsilon_V \Big]^T. \tag{3}$$

Observe that $\mathbf{v}_{j,i}$ is a $S_A \times 1$ vector whose elements are given by the elements of the sliding window which belong to a stationary region in the speech signal. The EV-neighborhood helps us to take into account a priori information about either the spread of the signal to be preserved or noise fluctuation to be suppressed.

http://dx.doi.org/10.13053/CyS-18-1-2014-023

2.1 Design of Proposed Robust Estimator

In conventional time-domain filtering an estimate of the *i*-th uncorrupted element of the signal $\{s_j(i)\}$ is a time-invariant function of elements of its sliding window $\mathbf{w}_{j,i}$, that is, $y_j(i) = EST(\mathbf{w}_{j,i})$. For instance, in the case of linear filtering the estimator is completely characterized by the impulse response of the filter. When the input signal is corrupted by a mixture of noise functions (see Eq. 1), a time-invariant filtering is unable to perform well. Therefore, a robust time-varying estimator is desirable [16, 15, 21]. Furthermore, since in some filtering algorithms the size of the sliding window is a constant, the algorithms can introduce undesirable blurring artifacts in nonstationary regions. We propose a locally adaptive rank-order algorithm in time domain which is capable to recover a noise-free speech signal employing a time-variant estimator over a locally adaptive neighborhood. We are interested in designing an optimum robust estimator to obtain an undistorted value $s_j(i)$ from the adaptive window $\mathbf{v}_{j,i}$, that is, $y_j(i) = EST(\mathbf{v}_{j,i})$. Suppose that a reference signal $x_j(i)$ is an estimate of the unavailable clean signal $s_j(i)$ from $\mathbf{v}_{j,i}$. The squared error between the estimated value $y_j(i)$ and the reference $x_j(i)$, is given by $e_j^2 = (y_j(i) - x_j(i))^2$. We want to design a robust estimator for minimizing the error. From the theory of robust estimation of location parameters [15, 22] three types of estimation of location parameters can be utilized to compute an estimate of the central element of the neighborhoods. They are the L-estimator based on linear combination of order statistics, the R-estimator derived from rank tests, and the M-estimator or the maximum likelihood estimator.

In this work we use the L-estimator which is one of the most popular robust estimators. The L-estimator over the EV-neighborhood is computed as follows:

$$y_j(i) = \mathbf{a}^T \mathbf{v}_{j,i}, \tag{4}$$

where $\mathbf{v}_{j,i}$ contains elements of the adaptive window of the size S_A and $\mathbf{a}$ is a $S_A \times 1$ vector of unknown weighting coefficients. In order to get an unbiased estimator the weighting coefficients must satisfy

$$\mathbf{a}^T \mathbf{u} = 1, \tag{5}$$

where $\mathbf{u}$ is a $S_A \times 1$ vector having only ones. Let $\mathbf{V}_{j,i}$ be a $S_A \times S_A$ diagonal matrix whose entries are elements of the vector $\mathbf{v}_{j,i}$, that is, $\mathbf{V}_{j,i} = DIAG(\mathbf{v}_{j,i})$. Denote a $S_A \times S_A$ diagonal matrix $\mathbf{R} = \mathbf{V}_{j,i}\mathbf{V}_{j,i}^*$, hence, we obtain $y_j^2(i) = \mathbf{a}^T \mathbf{R}\mathbf{a}$. The squared error can be rewritten as

$$\begin{aligned} e_j^2(i) &= x_j^2(i) + y_j^2(i) - 2x_j(i)y_j(i) \\ &= x_j^2(i) + \mathbf{a}^T\mathbf{R}\mathbf{a} - 2\mathbf{a}^T\mathbf{r}, \end{aligned} \tag{6}$$

where $\mathbf{r} = x_j(i)\mathbf{v}_{j,i}$ is a $S_A \times 1$ vector. The unknown vector $\mathbf{a}$ can be found by minimizing the squared error subject to satisfy Eq. 5. It can be seen that this is a constrained optimization problem which can be solved by the method of Lagrange multipliers minimizing the following objective function:

$$J(\mathbf{a}) = x_j^2(i) + \mathbf{a}^T\mathbf{R}\mathbf{a} - 2\mathbf{a}^T\mathbf{r} - 2(\mathbf{a}^T\mathbf{u} - 1)\lambda, \tag{7}$$

where λ is a constant. Since $\mathbf{R}$ is a non-negative matrix, the minimum value of Eq. 7 can be reached by solving

$$\frac{\partial}{\partial \mathbf{a}}\{J(\mathbf{a})\} = \mathbf{R}\mathbf{a} - \mathbf{r} - \mathbf{u}\lambda = 0. \tag{8}$$

The unknown coefficients are given by

$$\mathbf{a} = \mathbf{R}^{-1}(\mathbf{r} + \mathbf{u}\lambda). \tag{9}$$

From Eq. 5 we obtain λ, that is,

$$\lambda = (\mathbf{u}^T\mathbf{R}^{-1}\mathbf{u})^{-1}(1 - \mathbf{u}^T\mathbf{R}^{-1}\mathbf{r}). \tag{10}$$

Finally, substituting Eq. 10 into Eq. 9 and after some manipulations the vector $\mathbf{a}$ is given by

$$\mathbf{a} = \mathbf{R}^{-1}\left[\mathbf{r} + \frac{\mathbf{u}\left(1-\mathbf{u}^T\mathbf{R}^{-1}\mathbf{r}\right)}{\mathbf{u}^T\mathbf{R}^{-1}\mathbf{u}}\right]. \qquad (11)$$

Therefore, with the help of the L-estimator in Eq. 4, the uncorrupted signal can be recovered as follows:

$$y_j\left(i\right) = \mathbf{v}_{j,i}^T\mathbf{R}^{-1}\left[\mathbf{r} + \frac{\mathbf{u}\left(1-\mathbf{u}^T\mathbf{R}^{-1}\mathbf{r}\right)}{\mathbf{u}^T\mathbf{R}^{-1}\mathbf{u}}\right]. \qquad (12)$$

Now consider that the vectors $\mathbf{f}_j$ and $\mathbf{w}_{j,i}$ are corrupted with additive and impulsive noise. Suppose that the central element of the sliding window is located in a zone of abrupt transition of the speech signal and it is not an impulsive outlier. In this case, the signal of the sliding window $\mathbf{w}_{j,i}$ is nonstationary. On the other hand, the signal of the local EV-neighborhood $\mathbf{v}_{j,i}$ is a one-sided signal and it can be approximately considered as a stationary signal, that is,

$$\mathbf{v}_{j,i} = \mathbf{s}_{j,i} + \mathbf{b}_{j,i}. \qquad (13)$$

Observe that the signal in Eq. 13 is formed by samples of the clean signal and additive noise. Therefore, to obtain an estimate of the central signal element, conventional linear filtering uses all elements of the moving window, whereas rank-order filtering performs the estimation with statistically similar signal elements belonging to the adaptive neighborhood. The neighborhood does not contain impulsive outliers as well as signal elements with different statistical parameters (for instance, from other fragments of a signal to be processed). Various estimation strategies can be used to compute the reference signal $x_j\left(i\right)$. However, since the adaptive window $\mathbf{v}_{j,i}$ can be considered as a stationary signal formed by a clean signal corrupted by additive noise, an optimum estimator in the mean-squared-error sense is the Wiener filtering [6, 5]. It is well known that the frequency response of the Wiener filter is given by $H\left(\omega\right) = P_s\left(\omega\right) / \left(P_s\left(\omega\right) + P_b\left(\omega\right)\right)$, where $P_s\left(\omega\right)$ and $P_b\left(\omega\right)$ are the power spectral densities of a clean and noisy signals, respectively. For a short-time stationary segment with white clean and noise signals, the frequency response of the Wiener filter can be approximated by $H_{loc}\left(\omega\right) = \sigma_s^2 / \left(\sigma_s^2 + \sigma_b^2\right)$, where σ_s^2 and σ_b^2 are the variance of the clean and noisy speech segments, respectively. In time domain the impulse response of the Wiener filter is given by $h_{loc}\left(n\right) = \sigma_s^2 / \left(\sigma_s^2 + \sigma_b^2\right)\delta\left(n\right)$, where $\delta\left(n\right)$ is the Kronecker delta function. Under these considerations, the undistorted signal is estimated as

$$s_j\left(n\right) = n_z + \frac{\sigma_s^2}{\sigma_s^2 + \sigma_b^2}\left(v_z - n_z\right), \qquad (14)$$

where n_z is an estimate of the undistorted signal in a noisy speech (in silence periods of a speaker), v_z is an estimate of the undistorted signal in a voiced speech (in active periods of a speaker), and $G = \sigma_s^2 / \left(\sigma_s^2 + \sigma_b^2\right)$ is a local signal-to-noise ratio (SNR). If an input signal is corrupted with additive Gaussian noise, the maximum likelihood estimator is given by a sample mean. Therefore, a signal estimate in noisy speech can be computed as a sample mean over the neighborhood signal $n_z = \mu_{\mathbf{v}}$. The v_z value must be chosen to be close to the central value of the undistorted speech signal in the sliding window. Because of the signal inside of the adaptive window $\mathbf{v}_{j,i}$ is stationary, an approximation of the voice signal v_z can be obtained by smoothing the signal inside of the window, for instance, with the help of total variation denoising [24]. Note that this smoothing does not produce artifacts of temporal inertia. Let $\overline{\mathbf{v}_{j,i}}$ be a smoothed version of the adaptive window $\mathbf{v}_{j,i}$. The vector $\mathbf{r}$ in Eq. 12 can be computed by

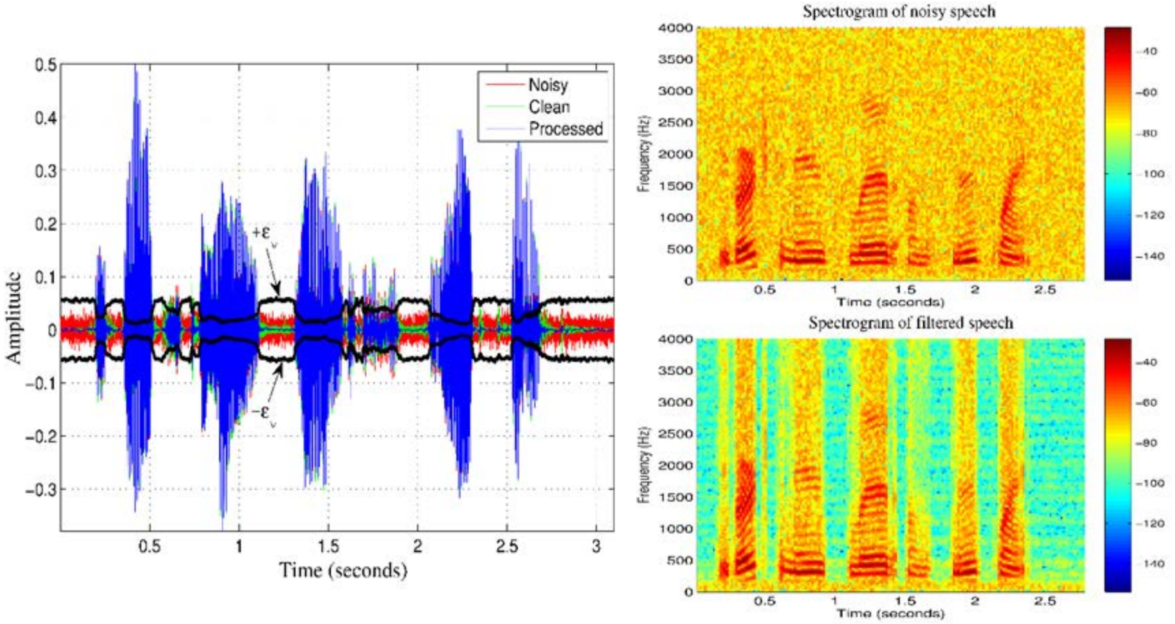


Fig. 2. Example of speech enhancement with proposed locally adaptive algorithm in 15 dB Gaussian noise

$$\mathbf{r} = \mathbf{v}_{j,i}\left[\mu_{\mathbf{v}} + \frac{\sigma_s^2}{\sigma_s^2+\sigma_b^2}\left(\overline{\mathbf{v}_{j,i}}(i)-\mu_{\mathbf{v}}\right)\right]. \qquad (15)$$

Recalling Eq. 14, we see that when the SNR is high, i.e. $\sigma_s^2 >> \sigma_b^2$, then $G \approx 1$ and the reference signal can be approximated by $x_j(i) = \overline{\mathbf{v}_{j,i}}(i)$. Otherwise, if the noise is much stronger than the signal, then $G \approx 0$ and the reference signal is given by $x_j(i) = \mu_{\mathbf{v}}$. Since the local variance of the adaptive window can be calculated as $\sigma_{\mathbf{v}}^2 = \sigma_s^2 + \sigma_b^2$, the local variance of the clean signal can be approximated with

$$\sigma_s^2 = MAX\left\{0,\left(\sigma_{\mathbf{v}}^2 - \sigma_b^2\right)\right\}. \qquad (16)$$

2.2 Design of the Proposed Robust Estimator

The proposed algorithm for robust speech processing consists of the following steps:

STEP 1: Read an initial input speech segment $\mathbf{n}_0$ with S elements assuming speaker's silence.

STEP 2: Read an input speech segment $\mathbf{f}_j$ (with N elements) to be processed (see Eq. 1).

STEP 3: Create a sliding window vector $\mathbf{w}_{j,i}$ around the noisy element $f_j(i)$ using Eq. 2.

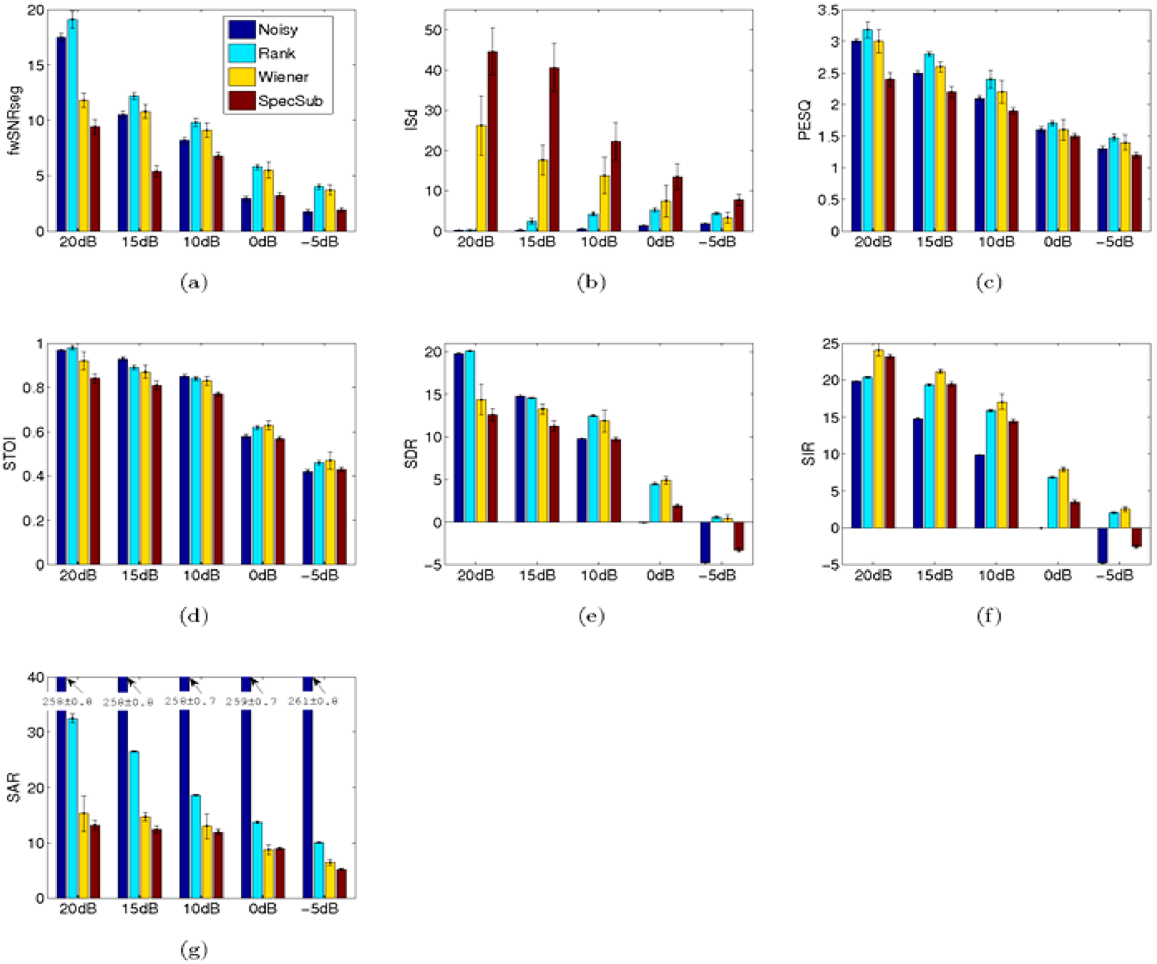
STEP 4: Calculate a local estimate of the SNR, as follows:

$$SNR_{j,i} = \frac{\mathbf{w}_{j,i}^T\mathbf{w}_{j,i}}{\mathbf{n}_0^T\mathbf{n}_0}. \qquad (17)$$

STEP 5: Calculate the ε_v -value, as follows:

$$\varepsilon_v(j,i) = \alpha_1\sigma_n\left[1 - \frac{1}{1+\left(SNR_{j,i}\right)^{-\alpha_2}}\right], \qquad (18)$$

where $\alpha_1 \geq 1$, and α_2 is within the range (0,1]. The parameters α_1 and α_2 help us to take into account a priori information about either the spread of the signal to be preserved or noise fluctuation to be suppressed. It is recommended in [25] to take $\alpha_1 = 1.5$. So, the interval of EV is given as $\pm 1.5\sigma_n$. The parameter α_2 of (0,1] is



(a) (b) (c)

(d) (e) (f)

(g)

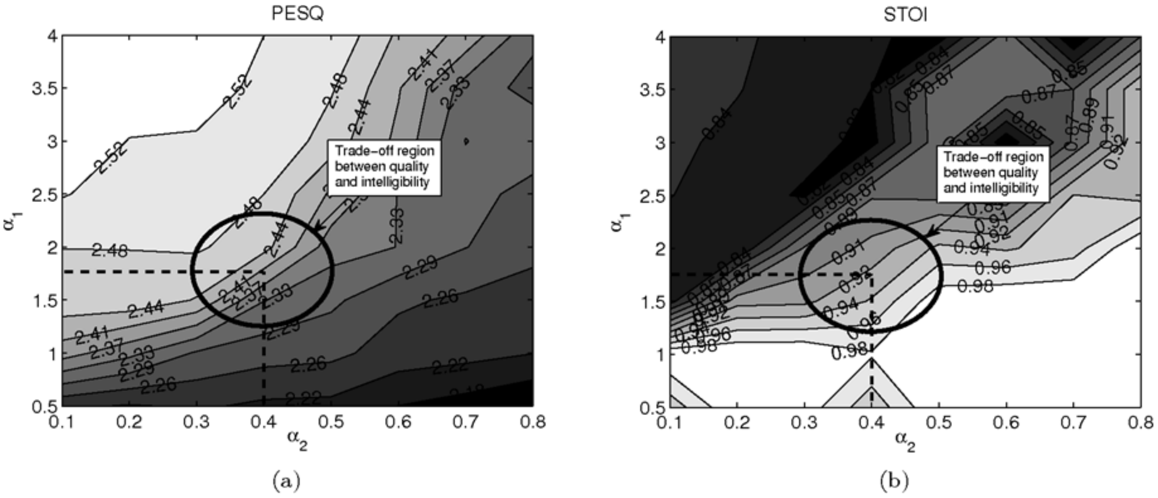
Fig. 3. Performance of speech enhancement algorithms with 95% confidence for processing speech corrupted with additive Gaussian noise at 20 dB, 15 dB, 10 dB, 0 dB and -5 dB SNRs. (a) fwSNRseg; (b) ISd; (c) PESQ; (d) STOI; (e) SDR; (f) SIR; (g) SAR

signal-dependent. When α_2 values are close to zero the changing of EV (dynamic range) as a function of the local SNR is small. In contrast, when α_2 values are close to unity the dynamic range of EV could be large. Actually, using these two parameters, a trade-off between noise suppression and introduction of artifacts to the processed speech signal can be achieved.

STEP 6: Construct an EV-neighborhood $\mathbf{v}_{j,i}$ from $\mathbf{w}_{j,i}$ with the help of Eq. 3 and Eq. 18.

At this point, the algorithm needs to identify either the central element of the sliding window is only corrupted by additive noise or the central element is an impulsive outlier. For each of these cases the algorithm performs in a different way. The outlier detection algorithm is as follows: if the size of the adaptive window $S_A = size\left(\mathbf{v}_{j,i}\right)$ is small, then EV-neighborhood contains only impulsive noise. We define a threshold parameter $L = \rho S$ for S_A, where ρ is the probability of occurrence of the impulsive noise and S is the

http://dx.doi.org/10.13053/CyS-18-1-2014-023



(a)

(b)

Fig. 4. Performance of the proposed algorithm versus α_1 and α_2 with respect to: (a) quality of processed speech (PESQ), (b) intelligibility (STOI)

size of $\mathbf{w}_{j,i}$. Thus, if $S_A < L$ then the outlier is detected.

STEP 7: Calculate the size of the adaptive window $S_A = size(\mathbf{v}_{j,i})$ and evaluate $S_A < L$. If the result is "false", compute vector $\mathbf{a}$ using Eq. 11 and go to STEP 8. If the result is "true", construct a new sliding window from the current one by excluding outlier elements of the adaptive neighborhood. Next, go to STEP 6.

STEP 8: Compute the output estimate $y_j(i)$ using Eq. 12.

STEP 9: Move the window $(i = i + 1)$ and evaluate if $i \le N$. If the result is "true" go to STEP 3. Otherwise, go to STEP 2.

In can be seen that the proposed algorithm is robust to nonstationary effects of a speech signal to be processed as well as to noise variation. When the noise is highly nonstationary [26], the local SNR in STEP 4 can be estimated with the help of a noise tracking algorithm [27].

3 Results

Computer experiments are carried out to evaluate and compare the performance of existing and proposed speech enhancement algorithms. We test the spectral subtraction [13], Wiener filtering [7], and the proposed algorithm for speech enhancement in environment of two different types of noise: stationary white Gaussian noise, and nonstationary street noise. Additionally, we test the algorithms in the framework of the CHiME challenge corpus [28]. This framework consists of processing of speech signals corrupted with noisy background conditions, collected from a real family living room. The considered spectral subtraction algorithm uses an adaptive gain averaging for reduction of musical noise [13]. In this approach, each frame is divided into smaller subframes to obtain a lower resolution spectrum. Therefore, the individual spectra in each subframe are subsequently averaged to obtain a lower-variance spectrum. On the other hand, the used Wiener algorithm is the one proposed by [7]. The algorithm is based on tracking a priori SNR estimate using the decision-directed approach [29]. For performance evaluation of tested algorithms several objective metrics are utilized: frequency weighted segmental signal-to-noise-ratio (fwSNRseg) [6], spectral-distance based Itakura Saito distance (ISd) [3], perceptual evaluation of speech quality (PESQ) [30], short-time objective intelligibility (STOI) [31], and composite metrics such as signal to distortion

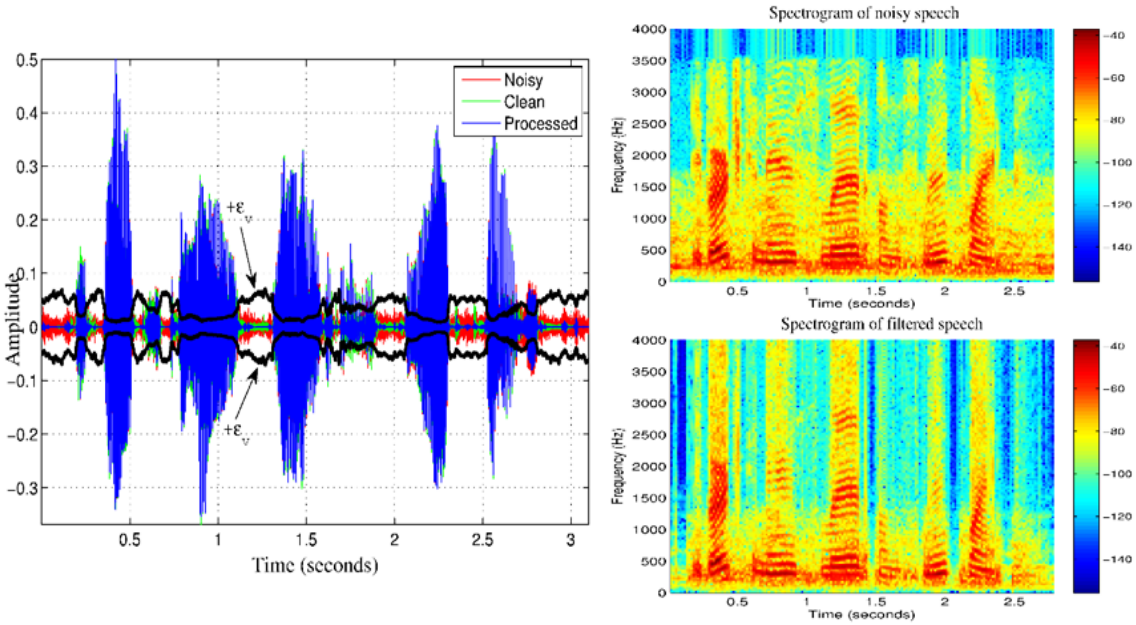
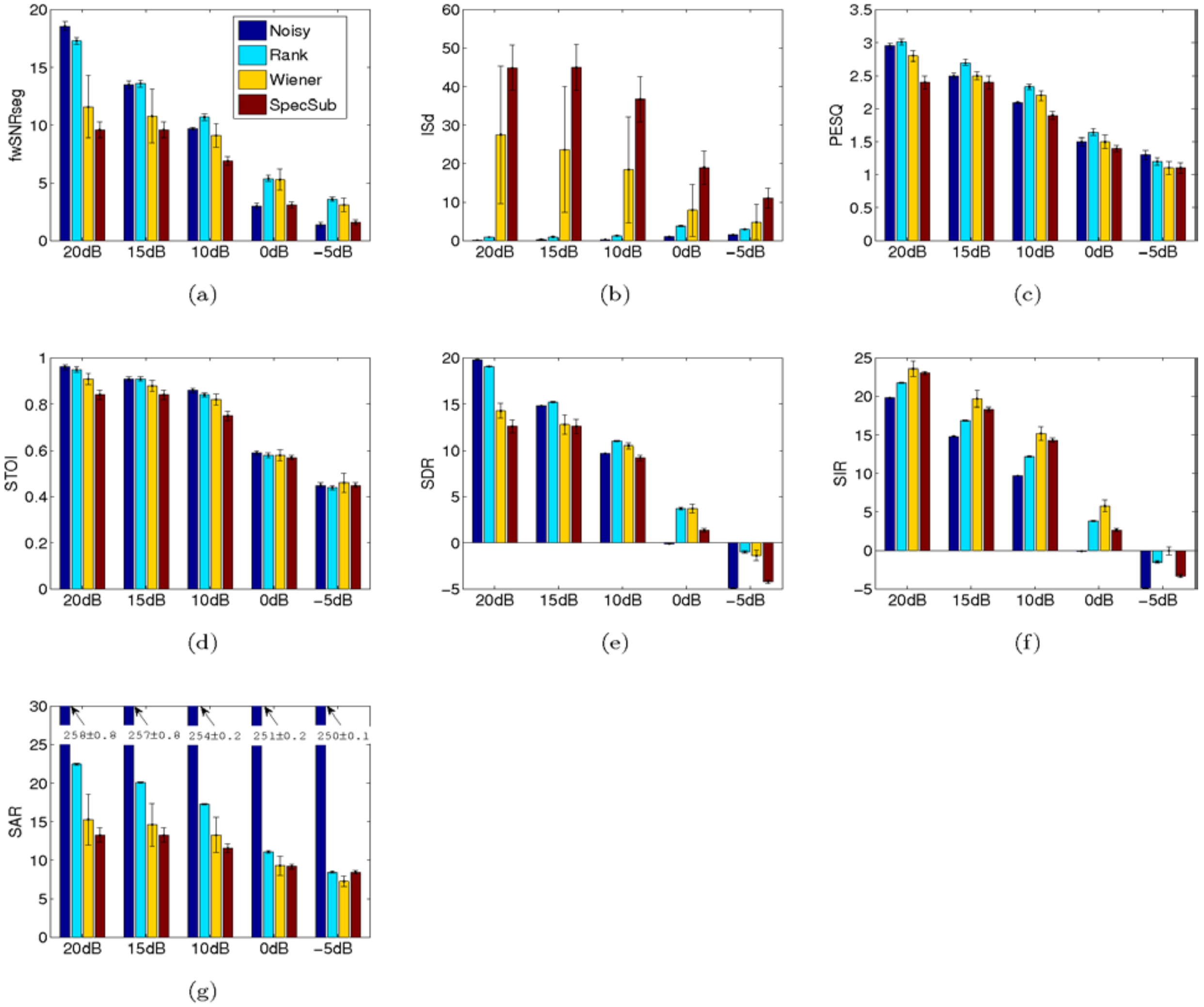


Fig. 5. Example of speech enhancement with proposed locally adaptive algorithm in 15 dB street noise

ratio (SDR), source to interference ratio (SIR), and source to artifacts ratio (SAR) [32]. Speech signals from the CHiME database [28] were used. The CHiME corpus contains 600 different sentences pronounced by male and female speakers in realistic noisy conditions. In our experiments, we will refer to the spectral subtraction, the Wiener filtering and the proposed algorithms as "SpecSub", "Wiener", and "Rank", respectively.

First we evaluate the performance of speech enhancement algorithms in stationary additive Gaussian noise while SNR is varied. The clean speech sentences from CHiME database were corrupted with stationary noise and processed with the tested algorithms. The window length and the initial silence period for the spectral subtraction and the Wiener filtering algorithms are 20 ms and 200 ms, respectively. The parameters for the proposed algorithm are $S = 65$ (7 ms), $\rho = 0.01$, $\alpha_1 = 1.5$, and $\alpha_2 = 0.5$. Fig. 2 shows examples of clean, noisy, and processed speech signals obtained with the proposed algorithm in Gaussian noise environment. It is interesting to note the change of $\pm\varepsilon_v$ values depending on local SNR of a noisy speech (see Eq. 18). Note that when α_2 values are close to zero we get a more aggressive filtering comparing with that of when α_2 values are close to unity. With 95% confidence the results in terms of the objective metrics for Gaussian noise are presented in Fig. 3. Observe that the proposed algorithm outperforms in majority of the cases the spectral subtraction and Wiener filtering algorithms in terms of speech quality given by the fwSNRseg, ISd, PESQ, and SDR metrics. In addition, we see that the proposed algorithm preserves better speech intelligibility that is characterized by the STOI, and introduces fewer artifacts according to the SAR than those of other tested algorithms. The spectral subtraction algorithm yields the worst results in terms of quality metrics and introduction of artifacts (see PESQ and SAR). However, this algorithm was the fastest one in our tests. The Wiener filtering yields acceptable results in terms of speech quality and intelligibility. Nevertheless, it introduces noticeable musical noise whereas the proposed algorithm does not do it. Actually, the amount of musical noise introduced by tested algorithms can be

http://dx.doi.org/10.13053/CyS-18-1-2014-023



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 6. Performance of speech enhancement algorithms with 95% confidence for processing speech corrupted with additive street noise at 20 dB, 15 dB, 10 dB, 0 dB, and -5 dB SNRs. (a) fwSNRseg; (b) ISd; (c) PESQ; (d) STOI; (e) SDR; (f) SIR; (g) SAR

characterized by the SAR. Observe from Fig. 3 that proposed algorithm yields the best SAR values in all performed tests.

Now the performance of the speech enhancement algorithms in framework of CHiME challenge corpus [28] is investigated. For the spectral subtraction and Wiener filtering algorithms the window length is 20 ms and the initial silence period (noise only) is 200 ms. The parameters for the proposed algorithm are given as follows: $S = 65$ (7 ms) and $\rho = 0.01$; α_1 and α_2 are taken as a trade-off between the PESQ and STOI metrics. This trade-off is found by evaluating the performance of the algorithm in terms of the PESQ and STOI while varying α_1 and α_2. The parameters are calculated for six different SNR values within the range of -6dB to 9dB using 120 speech signals from the development set of the CHiME corpus per each noise level.

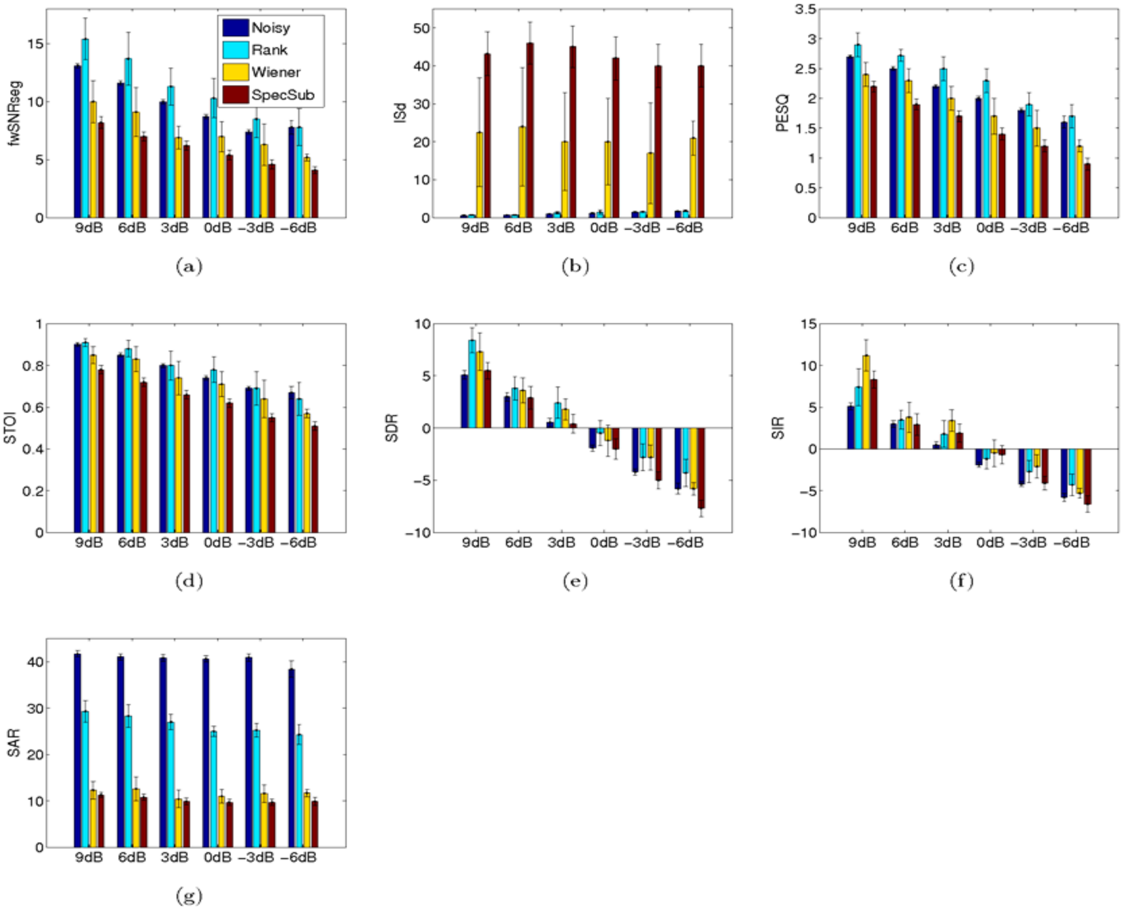
Next the performance of the speech enhancement algorithms using speech files from the development set of the CHiME corpus is tested. Instead, in this experiment a noise

http://dx.doi.org/10.13053/CyS-18-1-2014-023

tracking algorithm [27] is utilized to calculate local SNRs for the proposed algorithm. This algorithm is able to track fast changes in the local noise power spectrum from a noisy speech signal by using a data-driven recursive estimation of the noise power spectrum. With 95% confidence the results in terms of the performance metrics are presented in Fig. 7. It can be seen that the proposed algorithm yields a better performance in terms of quality measures such as fwSNRseg and ISd than that of the spectral subtraction and Wiener filtering algorithms. This means that the proposed algorithm introduces a lower level of spectral distortion to the processed speech comparing with the spectral subtraction and Wiener filtering algorithms. Furthermore, one can observe that the proposed algorithm is the only one which improves the PESQ value comparing with that of the noisy speech. This improvement is obtained without deteriorating the STOI performance. Note that the proposed algorithm also yields the best result with respect to introduction of artifacts (see SAR performance). Observe that the worst SAR value obtained with the proposed algorithm is about 25 dB whereas the best SAR value obtained with spectral subtraction and the Wiener filtering is about 12 dB. The Wiener filtering suppresses well the noise (the best performance in terms of the SIR) at the price of introducing a high level of artificial artifacts to the processed speech. The latter effect can be clearly seen from the performance of the Wiener filtering in terms of the SAR. The Wiener filtering yields good results in terms of noise reduction and background noise suppression, which are given by SIR and SDR measures. The spectral subtraction algorithm yields the worst results with respect to quality of speech perception and intelligibility. Additionally, this algorithm introduces many artifacts (see SAR performance). Finally, note that the proposed algorithm adaptively processes a noisy speech signal with a good trade-off between the speech quality and intelligibility.

Now, we discuss how to choose the parameters of the proposed algorithm: S , ρ , α_1 and α_2. The size S (length of the sliding window) must be chosen at least the double size of a pitch period of speech. The parameter ρ is the probability of occurrence of impulsive outliers. So, ρ can be estimated by calculating the number of impulsive outliers in silence periods of a speaker. To determine appropriate values for α_1 and α_2 , we evaluate the performance of the proposed algorithm with respect to the quality of processed speech (PESQ) and intelligibility (STOI) versus these two parameters. When speech is corrupted with 10 dB additive Gaussian noise, the results are shown in Fig. 4. It can be seen that the best speech quality is obtained for high values of α_1 and low values of α_2 . Note that a reasonable choice for the parameters α_1 and α_2 is given by those values that yield a trade-off between speech quality and intelligibility. According to Fig. 4, the trade-off corresponds to the values of α_1 and α_2 close to 1.7 and 0.4, respectively.

Next, the performance of the speech enhancement algorithms in nonstationary street noise is investigated. The clean speech sentences from the CHiME database were corrupted with a nonstationary noise. For the spectral subtraction and the Wiener filtering algorithms the window length is 20 ms and the initial silence period (noise only) is 200 ms. The parameters for the proposed algorithm are $S = 65$ (7 ms), $\rho = 0.01$, $\alpha_1 = 1.5$, and $\alpha_2 = 0.3$. Fig. 5 shows an example of clean, noisy, and processed speech signals with the proposed algorithm for a street noisy signal. Note that the proposed algorithm adapts well to nonstationary characteristics of the speech signals and to a nonstationary behavior of the background noise. With 95% confidence the results given in terms of the performance metrics for nonstationary street noise are presented in Fig. 6. We see that the proposed algorithm performs better in terms of the fwSNRseg, ISd, PESQ, and SDR than the Wiener filtering and spectral subtraction algorithms. The Wiener filtering yields good results in terms of speech quality and intelligibility. Actually, this filter yields a better performance with respect to noise reduction comparing with other tested algorithms (see the SIR metric). The spectral subtraction algorithm yields the worst results in terms of



(a) (b) (c) (d) (e) (f) (g)

Fig. 7. Performance of speech enhancement algorithms with 95% confidence for processing noisy speech sentences in the framework of CHiME challenge corpus. (a) fwSNRseg; (b) ISd; (c) PESQ; (d) STOI; (e) SDR; (f) SIR; (g) SAR

speech quality and intelligibility. This algorithm also introduces a noticeable musical noise that is characterized by the SAR. Note that the proposed algorithm preserves better speech intelligibility given by the STOI.

4 Conclusions

A locally adaptive algorithm for robust speech enhancement was presented. The algorithm is able to recover an undistorted signal from a noisy speech employing a time-variant estimator over a locally adaptive neighborhood. With the help of computer simulations, we showed that the proposed algorithm outperforms the spectral subtraction and Wiener filtering in terms of objective metrics. The proposed algorithm suppresses well additive noise and preserves high speech intelligibility by introducing lower levels of distortion.

References

1. **Oropeza-Rodriguez, J.L. (2006).** Algorithms and methods for the automatic speech recognition in spanish language using syllables. *Computación y Sistemas,* 9(3), 270–286.
2. **Ma, J. & Loizou, P.C. (2011).** SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech. *Speech Communication,* 53(3), 340–354.
3. **Hu, Y. & Loizou, P.C. (2008).** Evaluation of objective quality measures for speech

enhancement. *IEEE Transactions on Audio, Speech, and Language Processing,* 16(1), 229–238.

4. **Ma, J., Hu, Y., & Loizou, P.C. (2009).** Objective measures for predicting speech intelligibility in noisy conditons based on new band-importance functions*. The Journal of the Acoustical Society of America,* 125(5), 3387–3405.

5. **Vaseghi, S.V. (2008).** *Advanced digital signal processing and noise reduction* (4th ed.). Chichester, U.K.: J. Wiley & Sons.

6. **Loizou, P.C. (2007).** *Speech enhancement: theory and practice*. Boca Raton: CRC Press.

7. **Plapous, C., Marro, C., & Scalart, P. (2006).** Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing,* 14(6), 2098–2108.

8. **Scalart, P. & Filho, J.V. (1996).** Speech enhancement based on a priori signal to noise estimation. *IEEE International Conference on Acoustics, Speech and Signal Processing, Atlanta, GA, 2, 629–632.*

9. **Hansen, J.H.L. & Clements, M.A. (1991).** Constrained iterative speech enhancement with application to speech recognition. *IEEE Transactions on Signal Processing,* 39(4), 795–805.

10. **Sreenivas, T.V. & Kirnapure, P. (1996).** Codebook constrained Wiener filtering for speech enhancement. *IEEE Transactions on Speech and Audio Processing,* 4(5), 383–389.

11. **Boll, S. (1979).** Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics Speech and Signal Processing,* 27(2), 113–120.

12. **McAulay, R. & Malpass, M. (1980).** Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics Speech and Signal Processing,* 28(2), 137–145.

13. **Gustafsson, H., Nordholm, S.E., & Claesson, I. (2001).** Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Transactions on Speech and Audio Processing,* 9(8), 799–807.

14. **Hansler, E. (2008).** *Speech and audio processing in adverse environments*. New York: Springer.

15. **Astola, J. & Kuosmanen, P. (1997).** *Fundamentals of nonlinear digital filtering*. Boca Raton, Fla.: CRC Press.

16. **Yaroslavsky, L. (1996).** *Fundamentals of digital optics: digital signal processing in optics and holography.* Boston:Birkhäuser.

17. **Wang, S.S. & Lin, C.F. (1995).** Conditional trimmed mean filters and their applicacions for noise removal. *Signal Processing,* 43(1), 103–109.

18. **Coyle, E.J., Lin, J.H., & Gabbouj, M. (1989).** Optimal stack filtering and the estimation and structural approaches to image processing. *IEEE Transactions on Acoustics Speech and Signal Processing,* 37(12), 2037–2066.

19. **Gallagher, N.C. Jr. & Wise, G.L. (1981).** A theoretical analysis of the properties of median filters. *Transactions on Acoustics Speech and Signal Processing,* 29(6), 1136–1141.

20. **Arce, G.R. & McLoughlin, M.P. (1987).** Theoretical analysis of the max/median filter. *IEEE Transactions on Acoustics Speech and Signal Processing,* 35(1), 60–69.

21. **Kober, V.I., Mozerov, M.G., Alvarez-Borrego, J., & Ovseyevich, I.A. (2001).** Rank image processing using spatially adaptive neighborhoods. *Pattern Recognition and Image Analysis,*11(3), 542–552.

22. **Huber, P. J. (1981).** *Robust statistics*. New York: Wiley.

23. **Breithaupt, C., Gerkmann, T., & Martin, R. (2007).** Cepstral smoothing of spectral filter gains for speech enhancement without musical Noise. *IEEE Signal Processing Letters,* 14(12), 1036–1039.

24. **Karahanoglu, F.I., Bayram, I., & Van De-Ville, D. (2011).** A signal processing approach to generalized 1-D total variation. *IEEE Transactions on Signal Processing,* 59(11), 5265–5274.

25. **Pomalaza-Raez, C. & McGillem, C.D. (1984).** An adaptative, nonlinear edge-preserving filter. *IEEE Transactions on Acoustics Speech and Signal Processing,* 32(3), 571–576.

26. **Duan, Z., Mysore, G.J., & Smaragdis, P. (2012).** Speech Enhancement by Online Non-negative Spectrogram Decomposition in Non-stationary Noise Environments. *13th Annual Conference of the International Speech Communication Association* (*INTERSPEECH 2012),* Portland, Oregon, USA .

27. **Erkelens, J.S. & Heusdens, R. (2008).** Tracking of nonstationary noise based on data-driven recursive noise power Estimation. *IEEE Transactions on Audio, Speech and Languge Processing,* 16(6), 1112–1123.

28. **Christensen, H., Barker, J., Ma, N., & Green, P. (2010).** The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments. *11th Annual Conference of the International Speech Communication Association* (*INTERSPEECH 2010)*, Makuhari, Japan.
29. **Ephraim, Y. & Malah, D. (1985).** Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing,* 33(2), 443–445.
30. **ITU (2001).** *Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. *ITU-T Recommendation P.862*.
31. **Taal, C.H., Hendriks, R.C., Heusdens, R., & Jensen, J. (2011).** An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Acoustics, Speech and Signal Processing,* 19(7), 2125–2136.
32. **Vincent, E., Gribonval, R., & Févotte, C. (2006).** Performance measurement in blind audio source separation. *IEEE Transactions on Acoustics, Speech and Signal Processing,* 14(4), 1462–1469.
33. **Caballero-Morales S. O. & Trujillo-Romero F. (2013).** 3D Modeling of the Mexican sign language for a speech-to-sign languaje system. *Computación y Sistemas*, 17(4), 593–608.

Vitaly Kober received his Ph.D. and D.Sc. degrees in Image Processing from the Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, in 1992 and 2004, respectively. He is currently a professor at CICESE, Mexico. His research interests include signal and image processing, pattern recognition.

Victor Diaz Ramirez received his Ph.D. in Computer Science from CICESE, Mexico, in 2007. He is now a professor at the National Polytechnic Institute, CITEDI, Mexico. His research interests include signal and image processing, pattern recognition, and opto-digital correlators.

Yuma Sandoval Ibarra received her M.Sc. in Digital Systems from the National Polytechnic Institute, CITEDI, Mexico, in 2011. Now she is a Ph.D. student at the National Polytechnic Institute, CITEDI, Mexico. Her research interests include signal and speech processing.

Article received on 11/04/2013, accepted on 29/06/2013.

http://dx.doi.org/10.13053/CyS-18-1-2014-023