Speech Intelligibility Improvement in Noisy Environments for Near-End Listening Enhancement

Peyman Goli*
Department of Electronic and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran
p.goli@stu.nit.ac.ir
Mohammad Reza Karami-Mollaei
Department of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran
mkarami@nit.ac.ir

Received: 01/Jul/2015 Revised: 28/Feb/2016 Accepted: 02/Mar/2016

Abstract

A new speech intelligibility improvement method for near-end listening enhancement in noisy environments is proposed. This method improves speech intelligibility by optimizing energy correlation of one-third octave bands of clean speech and enhanced noisy speech without power increasing. The energy correlation is determined as a cost function based on frequency band gains of the clean speech. Interior-point algorithm which is an iterative procedure for the nonlinear optimization is used to determine the optimal points of the cost function because of nonlinearity and complexity of the energy correlation function. Two objective intelligibility measures, speech intelligibility index and short-time objective intelligibility measure, are employed to evaluate the noisy enhanced speech intelligibility. Furthermore, the speech intelligibility scores are compared with unprocessed speech and a baseline method under various noisy conditions. The results show large intelligibility improvements with the proposed method over the unprocessed noisy speech.

Keywords: Near-end Speech Enhancement; Intelligibility Improvement; Energy Correlation; Optimization Algorithms.

1. Introduction

Mobile phones often deliver speech output to listener in noisy environments. The background noise such as traffic or babble noise reduces speech intelligibility for the near-end listener. Several preprocessing algorithms have been proposed to improve speech intelligibility for the near-end listener in noisy environment. In speech improvement methods which focus on speech intelligibility enhancement for near-end listener in background noise, the far-end speech is considered as a clean speech with good intelligibility. As the far-end speech is played for near-end listener in noisy environment, its intelligibility is degraded by background noise; thus, these methods manipulate the clean speech (i.e., the far-end speech) before it is corrupted by the background noise to improve the intelligibility of noisy speech. Therefore, the clean speech and noise are available signals in the intelligibility enhancement methods and the aim is improvement of the audibility of degraded speech, as illustrated in Figure 1. Near-end speech intelligibility improvement methods have been classified to noise-independent and noise-dependent methods.

Noise-independent modification algorithms include detecting and boosting the features of speech that have an important role in speech perception. Charturong used hidden Markov model for detecting the consonant and transient regions [1], and Raset and Motlotle applied wavelet transform for extracting these regions in clean speech [2]. Demol et al. also used non-uniform time scaling to slow down the speech and redistributed available time between the vowels and consonants to emphasis on these regions [3]. In addition, Ekramul et al. presented a speech intelligibility improvement process in which speech is modified based on an inverse Wiener filter on the vowel and consonant regions [4].

Since the speech intelligibility in noisy environments usually depends on the noise conditions, noise-dependent algorithms may be applied to speech intelligibility enhancement in application scenario where the noise-statistics are available. Noise-dependent algorithms have been carried out using estimates of the noise signal. These methods are usually based on the signal to noise ratio (SNR) modification or optimization of an objective intelligibility measure. For example, Sauert and Enzner modified the local SNR of time-frequency cells based on the global SNR [5], and Tang and Cooke presented several strategies including the time and frequency segmentation and frequency selected boost to reach the global SNR [6]. Premananda and Uma also improved the near-end speech intelligibility focusing on the selective audible speech samples by considering the threshold of hearing and auditory properties of the human ear [7].

* Corresponding Author

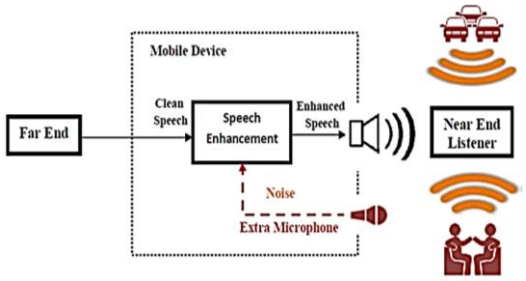


Fig. 1. Intelligibility enhancement of speech delivered in noisy environments. (Noise-independent/ Noise-dependent enhancement approach without/with extra microphone.)

Speech intelligibility enhancement based on maximization of speech intelligibility index (SII) measure [8], [9], and minimization of a perceptual distortion measure [10] are noise-dependent approaches which are considered as optimization algorithms. Tang and Cooke used a genetic algorithm-based optimization procedure, with glimpse proportion as the objective intelligibility metric to enhance the speech intelligibility in background noise [11], and Valentini-Botinhao et al. also increased the glimpse proportion measure by modifying mel-cepstral coefficients to improve the intelligibility of the synthetic speech in noise [12].

In the proposed algorithm, speech intelligibility is improved by optimizing the energy correlation between clean and noisy enhanced speeches in one-third octave frequency bands subjected to a power constraint. The fundamental idea behind the proposed method is that maximization of the cross correlation between speech degraded by noise and clean speech with good intelligibility would yield an improved speech intelligibility for the near-end listener. Hence, a cost function based on the energy correlation between the clean and noisy speeches is introduced to determine the cross correlation. Since the obtained cost function is complicated and nonlinear, the routine optimization methods (i.e., the derivative methods and Lagrange multiplier) cannot lead to an analytical solution. Therefore, an iterative algorithm is applied to optimize the cost function.

The paper is organized as follows: details of the proposed algorithm are described in three phases, namely, (1) preprocessing, (2) calculation and optimization of energy correlation function based on an iterative algorithm, and (3) estimation of statistical quantities. Finally, the objective intelligibility prediction results comparing the proposed algorithm with unprocessed speech and a baseline method are presented.

2. Proposed Speech Intelligibility Improvement Algorithm

To improve the speech intelligibility in background noise, the energy correlation function in one-third octave frequency bands within each time frame of the clean speech and speech degraded by noise, is determined and the correlation function is then optimized with a power speech constraint.

2.1 Preprocessing

Clean speech $x(n)$ and background noise $z(n)$ are available signals, and let $d(n)$ and $y(n) = d(n) + z(n)$ denote the enhanced speech and the speech degraded by noise, respectively. $x(n)$ and $z(n)$ are resampled by sample-rate of 10 kHz to capture a relevant frequency range for speech intelligibility [13]. Both signals are segmented into 50% overlapping, Hann-windowed frames with a length of 256 samples. The discrete Fourier transform (DFT) of each time frame is determined and a one-third octave band analysis is then performed by grouping DFT-bins. In total 15 one-third octave bands are used, where the lowest center frequency is set equal to 160 Hz and the highest octave band has a center-frequency equal to 4.06 kHz. Let $x(k, m)$ and $z(k, m)$ denote the DFT of the m^{th} frame of clean speech and noise in frequency index k, respectively. The energy of the j^{th} band in the m^{th} frame (i.e., $BF_{j,m}$) of clean speech, $X_{j,m}$, and noisy enhanced speech, $Y_{j,m}$, are calculated as follows,

$$X_{j,m} = \sum_{k=k_l(j)}^{k_h(j)} |x(k,m)|^2,$$

$$Y_{j,m} = \sum_{k=k_l(j)}^{k_h(j)} \left|G_{j,m}x(k,m) + z(k,m)\right|^2, \quad (1)$$

where $k_l(j)$ and $k_h(j)$ indicate the j^{th} one-third octave band edges and $|.|$ is the magnitude of the DFT. A real gain $G_{j,m} \in \mathcal{R}^+$, is applied to the $BF_{j,m}$ of the clean speech, to enhance the clean signal.

2.2 Calculation and Optimization of Correlation Function

The correlation coefficient is a statistical measure of the linear dependence between two random variables. Due to the energies $X_{j,m}$ and $Y_{j,m}$ are the random variables, the energy correlation function $\rho_{j,m}$ of the $BF_{j,m}$ of clean and noisy enhanced speeches is obtained as follows,

$$\rho_{j,m} = \frac{E(X_{j,m}Y_{j,m}) - E(X_{j,m})E(Y_{j,m})}{\sqrt{E(X_{j,m}^2) - E^2(X_{j,m})}\sqrt{E(Y_{j,m}^2) - E^2(Y_{j,m})}}, \quad (2)$$

where $E(.)$ indicates the expectation of the random variable. The numerator in Equation (2) determines the covariance of $X_{j,m}$ and $Y_{j,m}$, and the denominator is the product of theirs standard deviation. According to Equation (1) and the properties of the complex numbers, $Y_{j,m}$ would be rewritten as follows,

$$Y_{j,m} = G_{j,m}^2 X_{j,m} + G_{j,m} \sum_{k=k_l(j)}^{k_h(j)} x(k,m)z^*(k,m)$$

$$+ G_{j,m} \sum_{k=k_l(j)}^{k_h(j)} x^*(k,m)z(k,m) + Z_{j,m}, \quad (3)$$

where $.^*$ indicates the complex conjugate and $Z_{j,m}$ is the energy of the $BF_{j,m}$ of the noise. The energy correlation $\rho_{j,m}$ is obtained as a function of clean speech, noise and gain $G_{j,m}$ by substituting $Y_{j,m}$ in Equation (2).

Two obvious assumptions are considered to simplify the energy correlation: first, independency of clean speech and background noise, $E[x(n)z(n)] = E[x(n)]E[z(n)]$. Second, the zero expectation of the stochastic processes, clean speech and noise, $E[x(n)] = 0$ and $E[z(n)] = 0$.

According to Equations (2), (3) and the assumptions, the energy correlation function $\rho_{j,m}$ is simplified as follows,

$$\rho_{j,m} = \frac{\sigma_{X_{j,m}} G_{j,m}^2}{\sqrt{\sigma_{X_{j,m}}^2 G_{j,m}^4 + 4\Lambda_{j,m} G_{j,m}^2 + \sigma_{Z_{j,m}}^2}}, \quad (4)$$

Where

$$\Lambda_{j,m} = E\left\{\left[\mathrm{Re}\left(\sum_{k=k_l(j)}^{k_h(j)} x(k,m)z^*(k,m)\right)\right]^2\right\}, \quad (5)$$

$\mathrm{Re}(.)$ indicates the real part of the complex value. The statistical quantities $\sigma_{X_{j,m}}^2$ and $\sigma_{Z_{j,m}}^2$, which refer to the variance of the energies $X_{j,m}$ and $Z_{j,m}$, respectively, and $\Lambda_{j,m}$ are estimated from clean speech and noise. The energy correlation function $\rho_{j,m}$ is plotted in Figure 2 based on $G_{j,m}^2$, in the 10th frequency band for an SNR of –5 dB. The average of the energy correlation function of each frame $\sum_j \rho_{j,m}$ is maximized subjected to a power constraint to improve the speech intelligibility. The correlation function $\rho_{j,m}$ is concave in $G_{j,m}^2$, as illustrated in Figure 2. Hence, the sum of these concave functions, $\sum_j \rho_{j,m}$, is also concave. The constrained optimization problem can be formulated as follows,

$$\max_{G_{jm}} : \sum_{j=1}^{15} \frac{\sigma_{X_{j,m}} G_{j,m}^2}{\sqrt{\sigma_{X_{j,m}}^2 G_{j,m}^4 + 4\Lambda_{j,m} G_{j,m}^2 + \sigma_{Z_{j,m}}^2}};$$

$$\text{subject to} : \begin{cases} \sum_{j=1}^{15} G_{j,m}^2 X_{j,m} = P_m \\ G_{j,m}^2 \geq 0; \quad j = 1, \dots, 15 \end{cases}; \quad (6)$$

where $P_m = \sum_{j=1}^{15} X_{j,m}$ indicates the energy of the m^{th} frame of clean speech. The equality condition relates to the power constraint in the m^{th} frame, and the inequality condition satisfies the positive real gains $G_{j,m} \in \mathcal{R}^+$. The convexity is obtained by negation and the following Lagrangian cost-function characterizes the problem,

$$L = -\sum_{j=1}^{15} \frac{\sigma_{X_{j,m}} G_{j,m}^2}{\sqrt{\sigma_{X_{j,m}}^2 G_{j,m}^4 + 4\Lambda_{j,m} G_{j,m}^2 + \sigma_{Z_{j,m}}^2}} +$$

$$\lambda\left(\sum_{j=1}^{15} G_{j,m}^2 X_{j,m} - P_m\right) - \sum_{j=1}^{15} \omega_j G_{j,m}^2, \quad (7)$$

where λ and ω_j are Lagrangian multipliers related to the power constraint and inequality constraints in Equation (6), respectively. Since the objective function and constraints are differentiable, any point that satisfies the constraints in Equation (6) and the following conditions is guaranteed to optimize the problem [14].

1) $\mu_j \geq 0$; $\quad j = 1, \dots, 15,$

2) $\mu_j G_{j,m}^2 = 0; \quad j = 1, \dots, 15,$

3) $$\frac{\partial L}{\partial G_{j,m}^2} = -\frac{2\sigma_{X_{j,m}} \Lambda_{j,m} G_{j,m}^2 - \sigma_{X_{j,m}} \sigma_{Z_{j,m}}^2}{G_{j,m}^4 \sigma_{X_{j,m}}^2 + 4G_{j,m}^2 \Lambda_{j,m} + \sigma_{Z_{j,m}}^2}$$

$$+\lambda X_{j,m} - \omega_j = 0; \quad j = 1, \dots, 15, \quad (8)$$

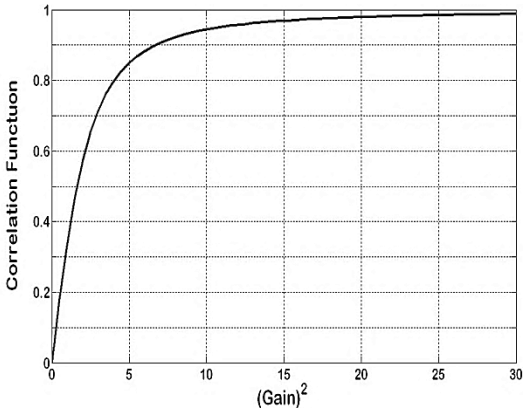


Fig. 2. Energy correlation function in the 10th band for an SNR of –5 dB based on $G_{j,m}^2$.

Because of complexity of the derivative of the Lagrangian cost-function, the optimization problem could not lead to an analytical formula. Thus, an iterative algorithm is applied to solve the optimization problem, which uses descent method to search the optimal point [14]. The Interior-point algorithm is an iterative method for non-linear optimization problem that consists of two steps. Step 1, inner loop, the Newton's method is applied to optimize the equality constrained problem. Step 2, outer loop, the Barrier method formulates the inequality constrained problem as an equality constrained problem to which Newton's method can be applied.

To formulate our constrained optimization problem as equality constrained problem, Equation (6) can be rewritten based on the Barrier logarithm as follows,

$$\min_{\breve{G}_{j,m}} : \tau f_0(\breve{G}_m) + \varphi(\breve{G}_m)$$

$$= -\tau \sum_{j=1}^{15} \frac{\sigma_{X_{j,m}} \breve{G}_{j,m}}{\sqrt{\sigma_{X_{j,m}}^2 \breve{G}_{j,m}^2 + 4\Lambda_{j,m} \breve{G}_{j,m} + \sigma_{Z_{j,m}}^2}} + \sum_{i=j}^{15} -\log(\breve{G}_{j,m}),$$

$$\text{subject to} : A\breve{G}_m = P_m, \quad (9)$$

where a simple variable change is used as $\breve{G}_{j,m} = G_{j,m}^2$. Vector $\breve{G}_m = [\breve{G}_{1,m}, \breve{G}_{2,m}, \dots, \breve{G}_{15,m}]^T$ consists of the gains of frequency bands in the m^{th} frame and vector $A = [X_{1,m}, X_{2,m}, \dots, X_{15,m}]$ refers to the energies of the clean speech bands. The equality $A\breve{G}_m = P_m$ shows the power constraint. $\varphi(\breve{G}_m) = -\sum_{j=1}^{15} \log(\breve{G}_{j,m})$ and $f_0(\breve{G}_m)$ are the Barrier logarithm and the negative energy correlation function based on $\breve{G}_m$, respectively. These functions are convex in $\breve{G}_m$. Parameter $\tau > 0$ sets the accuracy of the Barrier logarithm approximation. The optimization problem in the Equation (9) searches for the optimal vector which minimizes the cost function $f_0(\breve{G}_m)$ subjected to $A\breve{G}_m = P_m$ which satisfies power constraint and $\breve{G}_{j,m} > 0$ which is formulated in the Barrier logarithm. The optimization of the

energy correlation function using interior-point algorithm (Barrier and Newton's method) is summarized in Table I.

In interior-point algorithm shown in Table I, the inner loop minimizes the objective function $\tau f_0(\breve{G}_m) + \varphi(\breve{G}_m)$ (i.e., parameter τ is determined in outer loop) subjected to the power constraint using Newton's method. Hence, the Newton step $\Delta\breve{G}_m$ and decrement λ are calculated in the inner loop. In each iteration of the inner loop, the Newton step is added to $\breve{G}_m$ that was obtained in previous iteration, and the loop is run again until reaching inner stopping criterion, $\lambda^2/2 \le \acute{\epsilon}$ (i.e., for tolerance $\acute{\epsilon} > 0$). Backtracking line search is applied in the inner loop to determine the step size t used in gain updating at each Newton iteration. The optimal gain obtained by the inner loop is named central point, $\breve{G}_m^{*}(\tau)$, and delivered to the outer loop. In the outer loop, the central point $\breve{G}_m^{*}(\tau)$ is updated and then τ is increased by a factor $\mu > 1$. In other words, the central point $\breve{G}_m^{*}(\tau)$ is computed for a sequence of increasing values of τ until reaching the outer stopping criterion, $\tau \ge 15/\epsilon$ (i.e., for tolerance > 0), which guarantees the ϵ - suboptimal solution of the optimization problem.

Initialization of parameters of the interior-point algorithm largely affects the iterations of the inner and outer loops, and optimization accuracy. On one hand, excessive iterations of the loops lead to a large algorithmic delay; on the other hand, choosing large step sizes for reducing the iterations may results suboptimal points and reduces the accuracy. Therefore, the values of the parameters of the inner and outer loops are obtained from the experimental results, which provide the best performance of the interior-point algorithm in the optimization problem. These values yield the minimum iteration numbers and the maximum accuracy. The initialization of the parameters of the presented interior-point algorithm is shown in Table II.

Table 1. Optimization of energy correlation function using Barrier method (Newton's method).

Outer loop (Barrier Method)

Given feasible point $\breve{G}_m$, $\tau := \tau^{(0)} > 0$, $\mu > 1$, and tolerance $\epsilon > 0$.

Repeat:

1- Centering step

Inner loop (Newton's Method)

Starting point $\breve{G}_m$, tolerance $\acute{\epsilon} > 0$.

Repeat:

I. Compute the Newton step $\Delta\breve{G}_m$ and decrement λ:

$$\begin{bmatrix} \tau\nabla^2 f_0(\breve{G}_m) + \nabla^2\varphi(\breve{G}_m) & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} \Delta\breve{G}_m \\ \lambda \end{bmatrix} = \begin{bmatrix} -\tau\nabla f_0(\breve{G}_m) - \nabla\varphi(\breve{G}_m) \\ 0 \end{bmatrix}$$

$$\lambda^2 = [\tau\nabla f_0(\breve{G}_m) + \nabla\varphi(\breve{G}_m)]^T[\tau\nabla^2 f_0(\breve{G}_m) + \nabla^2\varphi(\breve{G}_m)]^{-1}[\tau\nabla f_0(\breve{G}_m) + \nabla\varphi(\breve{G}_m)]$$

II. Stopping criterion. Quit and $\breve{G}_m^{*}(\tau) = \breve{G}_m$ if $\lambda^2/2 \le \acute{\epsilon}$.

III. Line search by backtracking t:

Line search

i. Given descent direction $\Delta\breve{G}_m$, and $0 < \gamma < 0.5$, $0 < \theta < 1$

i. $t := 1$

i. While $\tau f_0(\breve{G}_m + t\Delta\breve{G}_m) + \varphi(\breve{G}_m + t\Delta\breve{G}_m) > \tau f_0(\breve{G}_m) + \varphi(\breve{G}_m) + \gamma t[\tau\nabla f_0(\breve{G}_m) + \nabla\varphi(\breve{G}_m)]^T\Delta\breve{G}_m$, $t := \theta t$

IV. Update: $\breve{G}_m := \breve{G}_m + t\Delta\breve{G}_m$

2- Updating $\breve{G}_m := \breve{G}_m^{+}(\tau)$. **(Starting point for next inner iteration)**

3- Stopping criterion. Quit if $\ge 15/\epsilon$ **.**

4- Increasing τ. $\tau := \mu\tau$. **END.**

Table 2. Initialization of the parameters of the inner and outer loops in the presented interior-point algorithm

Parameter Value	Loop
Feasible point $\breve{G}_m = [1,1,..,1]^T$	Outer
$\tau^{(0)} = 0.1$	Outer
$\mu = 2$	Outer
$\epsilon = 0.002$	Outer
$\acute{\epsilon} = 0.02$	Inner
$\gamma = 0.02$	Inner
$\theta = 0.25$	Inner

2.3 Estimation of Statistical Quantities

Given that the random variables $X_{j,m}$ and $Z_{j,m}$ are short-time stationary processes over the time frames, the statistical quantities could be estimated via time frame averaging [15],

$$\sigma^2_{X_{j,m}} = \frac{1}{N-1}\sum_{r=m-N+1}^{m}\left[X_{j,r} - E(X_{j,r})\right]^2,$$

$$E(X_{j,m}) = \frac{1}{N}\sum_{r=m-N+1}^{m} X_{j,r}, \quad (10)$$

and

$$\Lambda_{j,m} = \frac{1}{N-1}\sum_{r=m-N+1}^{m}\left[\mathrm{Re}\left(\sum_{k=k_l(j)}^{k_h(j)} x(k,r)z^*(k,r)\right)\right]^2, \quad (11)$$

where N denotes the number of successive frames in which the quantities are estimated. The best results are obtained in $N = 30$. Such as Equation (10), similar estimation hold for $\sigma^2_{Z_{j,m}}$. In practice, a simple noise-tracker algorithm [16] could be applied to estimate the power of noise signal.

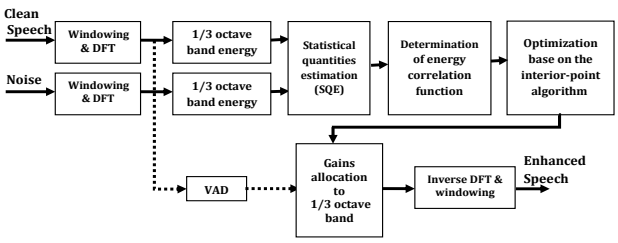


Fig. 3. Block diagram of the proposed algorithm

A simple smoother is then applied to the statistical quantities to prevent high changes which may negatively affect the estimation,

$$\hat{\sigma}^2_{X_{j,m}} = \alpha\hat{\sigma}^2_{X_{j,m-1}} + (\alpha - 1)\sigma^2_{X_{j,m}}, \quad (12)$$

where $\alpha = 0.9\,6$ leads to best results and similar smoother is applied to $\sigma^2_{Z_{j,m}}$ and $\Lambda_{j,m}$.

3. Proposed Algorithm Implementation

A simple voice activity detector (VAD) is used in the proposed algorithm, as illustrated in Figure 3. The VAD block selects the speech frames whose power is greater than $P_{max} - K$ for the process. P_{max} is the maximum power of the received frames of clean speech in dB and the constant value of K is selected 25 dB in practice. Clean speech, enhanced speech by proposed method and

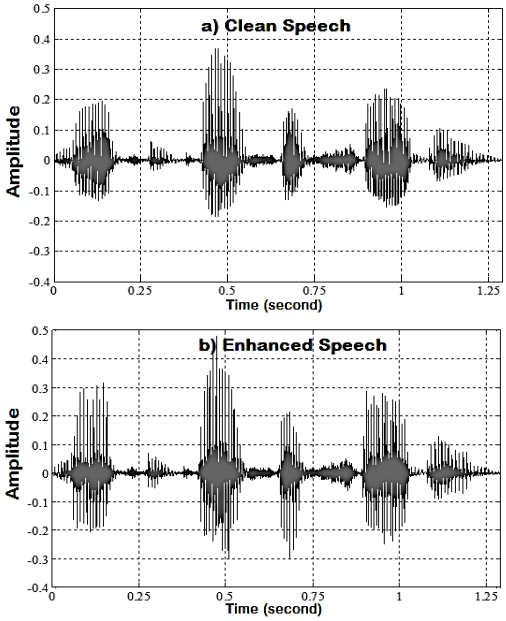
iteration number of the interior-point algorithm within each time frame in white noise for an SNR of –5 dB are shown in Figure 4. The iteration number is equal to zero in silent frames as illustrated in Figure 4, since the VAD block prevents to manipulate these frames. Also, the iteration range of the interior-point algorithm in speech-active frames is 1 to 167 with a total of 5302 iterations.

The enhanced speech is corrupted by white noise and babble noise to evaluate the performance of the proposed algorithm. The spectral power of white noise is uniformly distributed over the frequencies. This important property makes white noise appropriate to evaluate the performance of the speech enhancement methods in frequency domain; however, white noise is not an environmental noise. Babble noise, the noise of the public places, is a common environment noise that may destroy speech intelligibility in wireless communications.

The effect of the proposed algorithm on the power of one-third octave bands in clean and enhanced speech for white and babble noises with an SNR of –5 dB is shown in Figure 5. The energy is usually distributed from the frequency bands in which the clean speech power is greater than the noise power to other frequency bands, as illustrated in Figure 5. This transmission is such that the average of the energy correlation can be maximized. In other words, this method spreads the energy between the bands and makes a dense spectral power density.

4. Performance Evaluation

To evaluate the performance of the proposed algorithm, the clean and enhanced speech is degraded by white, babble, factory, and traffic noises at the SNRs of –20, –15, –10, –5 and 0 dB from the NOISEX-92 database.



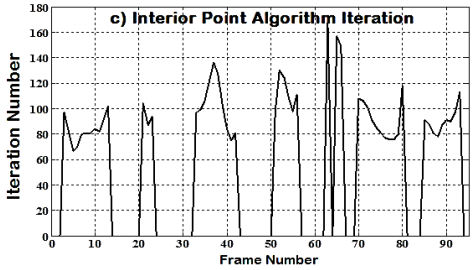


Fig. 4. The time domain plots of (a) clean speech and (b) enhanced speech with an SNR of –5 dB, and (c) iteration number of the interior-point algorithm within each frame.

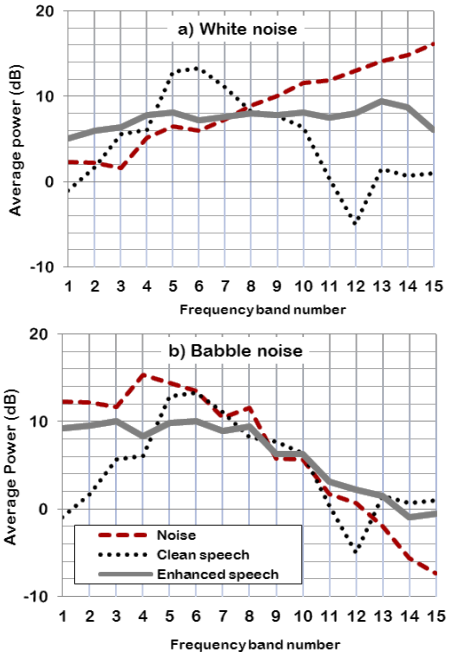


Fig. 5. Avrage power of one-third octave bands of the N=30 successive frams for noise, clean speech and enhaced speech in (a) white noise and (b) babble noise with an SNR of –5 dB.

In total, 20 random sentences from male speakers are used from the TIMIT database. The duration of each sentence is almost between 3 s and 5 s. A comparison is made with the unprocessed noisy speech and a reference method proposed by Taal *et al.* [10]. This method, which is similar to the proposed method, improves speech intelligibility by optimizing a cost function. Taal *et al.* optimally redistribute the speech energy over time-frequency cells according to a perceptual distortion measure. The baseline algorithm minimizes the mentioned objective measure by using auditory filter bank with a power constraint over all speech-active frames to improve the speech intelligibility.

The current study employs two objective intelligibility measures to predict the intelligibility of the noisy enhanced speech. The objective measures enable rapid feedback on a range of speech intelligibility enhancement methods. First measure is speech intelligibility index (SII), which is based on weighted SNRs. In the one-third octave band procedure provided by ANSI [17], the SII measure is computed by dividing the spectrum of clean speech and noise into one-third octave frequency bands and estimating the weighted average of the SNRs in each

band. The SNRs are weighted by band importance functions which differ across speech materials. The output of the SII measure is a scalar number between 0 and 1, which predicts the speech intelligibility in background noise. Second method is short-time objective intelligibility (STOI) measure, which is based on a correlation coefficient between the temporal envelopes (i.e., the root of energy of one-third octave frequency bands) of the clean and degraded speech in overlapping segments [18]. To calculate the STOI measure, the clean and degraded speech are decomposed into DFT-based. Then, short-time temporal envelope segments of the clean and degraded speech are compared by means of a correlation coefficient after normalizing and clipping. The STOI score is then obtained by averaging these short-time intelligibility measures over the speech signal. This measure provides a score in the range of 0 to 1, which refers to the speech intelligibility of degraded speech. Both measures can predict the intelligibility of noisy speech in various speech degradations.

The proposed method provides the best intelligibility scores in the STOI measure in comparison with the unprocessed noisy speech and the baseline method proposed by Taal *et al.* for all noisy conditions, as illustrated in Figure 6. The results also show that the baseline method led to a decrease in model intelligibility based on STOI at low SNRs in all noises except babble noise in which it provides a large increase in comparison with the unprocessed noisy speech in all SNRs.

Figure 7 presents the SII measure intelligibility scores. The results show intelligibility improvement with the proposed algorithm in comparison with the unprocessed noisy speech and the baseline method at low SNRs in the SII measure for white, babble, and traffic noises. However, the baseline method obtains better SII scores than our algorithm in factory noise at all SNRs. The baseline method also provides better intelligibility scores at –5 and 0 dB than the proposed method in all maskers, as illustrated in Figure 7.

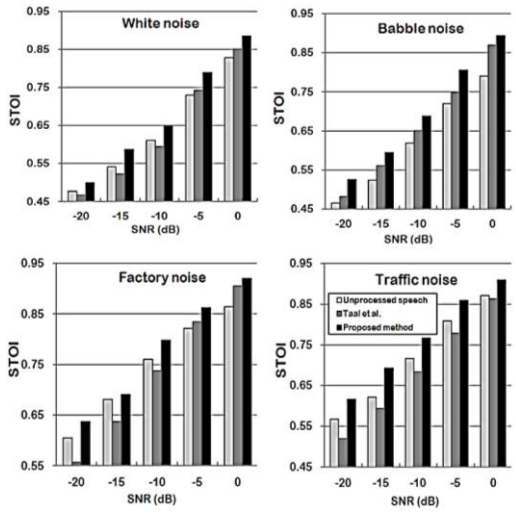


Fig. 6. STOI intelligibility predictions for the proposed method, unprocessed noisy speech, and the baseline method by Taal et al. for white, babble, factory, and traffic noises at the SNRs of –20, –15, –10, –5 and 0 dB.

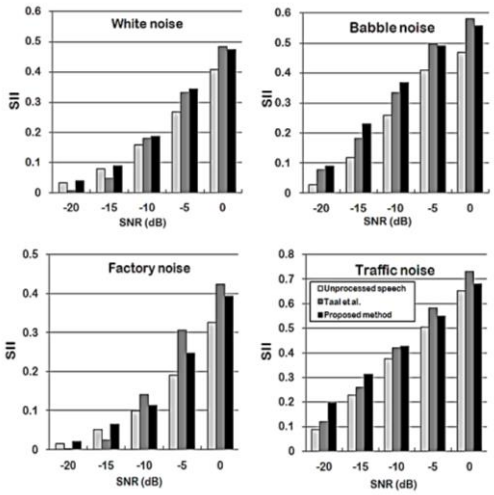


Fig. 7. SII intelligibility predictions for the proposed method, unprocessed noisy speech, and the baseline method by Taal et al. for white, babble, factory, and traffic noises at the SNRs of –20, –15, –10, –5 and 0 dB.

The significant intelligibility scores that the proposed method obtained in the STOI predictor are expected, since our method maximizes the energy correlation between clean speech and noisy enhanced speech, and also the STOI measure is based on the mean cross-correlations of the root of energy in frequency bands between these signals. Thereby, both algorithms are based on the correlation between clean speech and noisy speech. The proposed method improves speech intelligibility based on the correlation and STOI measures it to predict the speech intelligibility. Therefore, it can be stated that the proposed method maximizes an intelligibility cost function according to the STOI measure.

5. Conclusions

A new speech intelligibility improvement algorithm is proposed to enhance the speech intelligibility for the near-end listener in noisy environments without increasing the speech energy. This was performed by maximizing the energy cross correlation of the one-third octave bands between clean speech and enhanced noisy speech with a power constraint. The interior-point algorithm, an iterative algorithm for nonlinear optimization, is applied to solve the optimization problem, because of the nonlinearity and complexity of the cost function. The speech energy is redistributed over the frequency bands of clean speech according to the optimization of the energy correlation between clean and noisy speech. Two objective intelligibility predictors, the STOI and SII measures, are employed for scoring the intelligibility of the noisy enhanced speech under various noisy conditions to evaluate the performance of the proposed method. The results show significant intelligibility improvement with the proposed algorithm in comparison with the unprocessed noisy speech. As the current work is a frame-based speech enhancement method that maximizes the

cost function within each time frame, the proposed method can be appropriate for online processing. However, the iterative optimization algorithm used in the optimization problem is not appropriate for online processing due to the algorithmic delay produced with the loop iterations. Therefore, the iterative algorithm would be replaced with an alternative method with an insignificant algorithmic delay, in future works.

Acknowledgments

The research is supported by Khavaran Institute of Higher Education (KHI).

References

[1] C. Tantibundhit, J. R. Boston, C. C. Li, D. J. Durrant, S. Shaiman, K. Kovacyk, and A. El-Jaroudi. "Speech enhancement using transient speech components," in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2006, pp. 1–4.

[2] D. M. Rasetshwane. Enhancement of speech intelligibility using speech transients extracted by a wavelet packet-based real-time algorithm. Ph.D. Thesis, University of Pittsburgh, 2009.

[3] D. Mile, W. Verhelst, K. Struyve, and P. Verhoeve. "Efficient non-uniform time-scaling of speech with WSOLA," in: Proc. SPECOM, 2005, pp. 163–166.

[4] M. E. Hamid, S. Das, K. Hirose, and M. K. I. Molla. "Speech enhancement using EMD-based adaptive soft-thresholding EMDADT," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 5, no. 2, pp. 1–16, 2012.

[5] B. Sauert, G. Enzner, and P. Vary. "Near-end listening enhancement with strict loudspeaker output power constraining," in: Proc. IWAENC, 2006. pp. 1–4.

[6] Y. Tang and M. Cooke. "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in: Proc. Interspeech, 2011, pp. 345–348.

[7] B. S. Premananda and B. V. Uma. "Selective Frequency Enhancement of Speech Signal for Intelligibility Improvement in Presence of Near-end Noise," Computer Science, Elsevier, vol. 49, pp. 244–252, 2015.

[8] B. Sauert and P. Vary. "Near-end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in: Proc. 18th European Signal Processing Conference, EUSIPCO, 2010, pp. 1919–1923.

[9] C.H. Taal, J. Jensen, and A. Leijon. "On optimal linear filtering of speech for near-end listening enhancement," Signal Processing Letters, IEEE, vol. 20, no. 3, pp. 225–228, 2013.

[10] C. H. Taal, R. C. Hendriks, and R. Heusdens. "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," Computer Speech & Language, Elsevier, vol. 28, no. 4, pp. 858–872, 2014.

[11] Y. Tang and M. Cooke. "Optimized spectral weightings for noise-dependent speech intelligibility enhancement," in: Proc. Interspeech, 2012, pp. 955–958.

[12] C. Valentini-Botinhao, J. Yamagishi, S. King, and R. Maia, "Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion," Computer Speech and Language, Elsevier, vol. 28, no. 2, pp. 665–686, 2014.

[13] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," Journal of Acoustical Society of America, vol. 19, no. 1, pp. 90–119, 1947.

[14] S. Boyd and L. Vandenberghe. Convex optimization. Cambridge university press, 2004, pp. 561–615.

[15] P. Vary and R. Martin. Digital speech transmission: Enhancement, coding and error concealment. John Wiley & Sons, 2006, pp. 143–150.

[16] T. Gerkmann and R. C. Hendriks. "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay." IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 4, pp. 1383 – 1393, 2012.

[17] ANSI. S3.5-1997 Methods for Calculation of the Speech Intelligibility Index. New York: American National Standards Institute, 1997, pp. 90–119.

[18] C. H. Taal, R. Hendriks, R. Heusdens, and J. Jensen. "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 7, pp. 2125 – 2136, 2011.

Peyman Goli received the B.Sc and M.Sc degrees in Electronic Engineering from Noshirvani University of Technology, Babol, Iran, in 2002 and 2005, respectively. He is currently pursuing the Ph.D degree at the Electronic Engineering (Signal Processing), Noshirvani University of Technology, Babol, Iran, under the supervision of Dr. Mohammad Reza Karami-Mollaei. His main research topic is speech intelligibility enhancement in noisy environments. Other research interests include acoustic echo cancelation, speech noise cancelation and signal processing in the field of digital audio.

Mohammad Reza Karami-Mollaei received the B.Sc in Electrical and Electronic Engineering in 1992, M.Sc of signal processing in 1994, and Ph.D in 1998 in Biomedical Engineering from I.N.P.L d'Nancy of France. He is now an associate professor with the Department of Electrical and Computer Engineering, Babol University of Technology. Since 1998 his research is in signal and speech processing. He published more than 100 articles in journals and conferences. He teaches Digital Signal, Biomedical and speech processing at university. His research interests include Speech, Image and signal processing.