PAPER *Special Issue on Speech Information Processing*

Grey Filtering and Its Application to Speech Enhancement

Cheng-Hsiung HSIEH[†], ***Regular Member***

SUMMARY In this paper, a grey filtering approach based on GM(1,1) model is proposed. Then the grey filtering is applied to speech enhancement. The fundamental idea in the proposed grey filtering is to relate estimation error of GM(1,1) model to additive noise. The simulation results indicate that the additive noise can be estimated accurately by the proposed grey filtering approach with an appropriate scaling factor. Note that the spectral subtraction approach to speech enhancement is heavily dependent on the accuracy of statistics of additive noise and that the grey filtering is able to estimate additive noise appropriately. A magnitude spectral subtraction (MSS) approach for speech enhancement is proposed where the mechanism to determine the non-speech and speech portions is not required. Two examples are provided to justify the proposed MSS approach based on grey filtering. The simulation results show that the objective of speech enhancement has been achieved by the proposed MSS approach. Besides, the proposed MSS approach is compared with HFR-based approach in [4] and ZP approach in [5]. Simulation results indicate that in most of cases HFR-based and ZP approaches outperform the proposed MSS approach in SNR_{imp}. However, the proposed MSS approach has better subjective listening quality than HFR-based and ZP approaches..

key words: *grey filtering, GM(1,1) model, additive noise, estimation error, speech enhancement, spectral subtraction*

1. Introduction

The purpose of filtering is to recover signal component from noisy observations [1]. Filtering is required in many engineering applications. One example is the speech enhancement. Assume that the signal model is the additive signal model, which is expressed as $x(k) = s(k) + n(k)$ where $x(k)$, $s(k)$, and $n(k)$ are noisy speech, clean speech, and the additive noise, respectively. The objective of speech enhancement is to recover $s(k)$ from noisy speech $x(k)$. Note that filtering $s(k)$ out of $x(k)$ is equivalent to the estimation of additive noise $n(k)$. Therefore better performance of speech enhancement results from appropriate noise estimation. Basically, the speech enhancement consists of two stages: noise estimation and noise removal. Up to present, several noise estimation approaches have been reported. Some of representative approaches are as follows. By a mechanism to determine non-speech and speech portions in $x(k)$, additive noise is estimated during non-speech period in [2]. Note that the spectrum above speech frequency component comes from $n(k)$ if it is white noise. In [3], the spectral component of white noise is estimated through linear prediction coefficients while higher sampling rate is used for spectral estimation of $n(k)$ in [4]. By signal insertion in the transmitted speech signal, in [5] the contaminated inserted signals are used to estimate noise. Since additive noise is random, it is appropriate to deal with $n(k)$ in a statistical way. Therefore, statistics of $n(k)$ is sufficient in many practical applications of speech enhancement. When statistics of noise are estimated, a noise removal technique is applied. The noise removal approach can be Weiner filtering as in [2], Kalman filtering as in [3], or a popular approach called spectral subtraction as in [4] and [5].

In this paper, we proposed a grey filtering approach based on GM(1,1) model [6] which stands for the first-order grey model with one variable. Then the proposed grey filtering approach is applied to speech enhancement whose noise removal technique is based on magnitude spectral subtraction (MSS) [7]. This paper is motivated by the following observations. The estimation error of GM(1,1) model is zero for a constant signal and approximately zero for random signal when additive noise is absent. When additive noise is present, both signals have non-zero estimation error. These results will be shown in Sect. 2.2. By the observations, it implies that estimation error of GM(1,1) model can be related to additive noise. Furthermore, the speech signal generally consists of two parts: non-speech and speech. The non-speech portion can be considered as constant signal while speech portion as random signal. Consequently, there is a hope to estimate additive noise in noisy speech through estimation error of GM(1,1) model and therefore speech enhancement by spectral subtraction may be possible.

This paper is organized as follows. In Sect. 2, a brief review of GM(1,1) model [6] is given and grey filtering or noise estimation based on GM(1,1) model is described and demonstrated as well. In Sect. 3, the application of grey filtering to MSS [7] for speech enhancement is proposed and described. Then simulation results are provided to justify the proposed MSS approach in Sect. 4 where comparisons with approaches in [4] and [5] are made as well. Finally, conclusive remarks are made in Sect. 5.

Manuscript received July 3, 2002.

Manuscript revised November 18, 2002.

[†] The author is with the Department of Electronic Engineering, Chien Kuo Institute of Technology, Changhua, Taiwan, R.O.C.

2. Grey Filtering Based on GM(1,1) Model

In this section, a brief review of GM(1,1) model is given first and then the grey filtering approach to noise estimation is described and demonstrated.

2.1 GM(1,1) Model

The GM(1,1) modeling process is described in the following. For details, one may consult [6]. Given data sequence $\{x(k)$, for $1 \leq k \leq K\}$, a new data sequence $x^{(1)}(k)$ is found by 1-AGO (first-order accumulated generating operation) as

$$x^{(1)}(k) = \sum_{n=1}^{k} x(n) \tag{1}$$

for $1 \leq k \leq K$, where $x^{(1)}(1) = x(1)$. To be effective in GM(1,1) modeling, $x(k)$ needs to meet two conditions: (i) data is of same sign, and (ii) the ratio between adjacent data in $x(k)$ should be less than 10. From (1), it is obvious that the original data $x(k)$ can be easily recovered from $x^{(1)}(k)$ as

$$x(k) = x^{(1)}(k) - x^{(1)}(k-1) \tag{2}$$

for $2 \leq k \leq K$. This operation is called 1-IAGO (first-order inverse accumulated generating operation).

By sequences $x(k)$ and $x^{(1)}(k)$, a grey difference equation is formed as

$$x(k) + az^{(1)}(k) = b \tag{3}$$

where

$$z^{(1)}(k) = 0.5(x^{(1)}(k) + x^{(1)}(k-1)) \tag{4}$$

for $2 \leq k \leq K$, and parameters a and b are called developing coefficient and grey input, respectively. From (3), parameters a and b can be obtained as

$$\begin{bmatrix} a \\ b \end{bmatrix} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y} \tag{5}$$

where

$$\mathbf{B} = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(K) & 1 \end{bmatrix} \tag{6}$$

and

$$\mathbf{y} = \begin{bmatrix} x(2) \\ x(3) \\ \vdots \\ x(K) \end{bmatrix} \tag{7}$$

Basically, the difference equation in (3) is attempted to approximate differential equation in the GM(1,1) modeling. Therefore, to find the solution of $x^{(1)}(k)$ in (3), we utilize its associated differential equation, which has the following form

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = b \tag{8}$$

It can be easily shown that the solution for $x^{(1)}(t)$ in (8) is

$$x^{(1)}(t) = ce^{-at} + \frac{b}{a} \tag{9}$$

where c, by the initial condition $x^{(1)}(t_0) = x(t_0)$, can be found as

$$c = \left(x(t_0) - \frac{b}{a}\right) e^{at_0} \tag{10}$$

Therefore, the solution for $x^{(1)}(t)$ is given as

$$x^{(1)}(t) = \left(x(t_0) - \frac{b}{a}\right) e^{-a(t-t_0)} + \frac{b}{a} \tag{11}$$

Letting $t_0 = 1$ and $t = k$, we have the solution of $x^{(1)}(k)$ as follows.

$$x^{(1)}(k) = \left(x(1) - \frac{b}{a}\right) e^{-a(k-1)} + \frac{b}{a} \tag{12}$$

where parameters a and b are found in (5). By 1-IAGO, the estimate of $x(k)$, $\hat{x}(k)$, is obtained as

$$\hat{x}(k) = x^{(1)}(k) - x^{(1)}(k-1) \tag{13}$$

where $\hat{x}(1) = x^{(1)}(1) = x(1)$. The estimation error for $x(k)$ is given as

$$e(k) = x(k) - \hat{x}(k) \tag{14}$$

which will be used to estimate additive noise later in Sect. 2.2.

To sum up, the GM(1,1) modeling process consists of three steps. First, find parameter a and b by (5). Second, use (12) to estimate $x^{(1)}(k)$. Finally, find $\hat{x}(k)$ through (13). It should be noticed that the minimum number of data samples in GM(1,1) modeling is as few as four samples, i.e. $K = 4$.

2.2 Grey Filtering and Noise Estimation

The proposed grey filtering approach based on GM(1,1) model is described here. Assume the available noisy signal $x(k)$ satisfies Conditions (i) and (ii) in Sect. 2.1 and has the additive signal model $x(k) = s(k) + n(k)$ where $s(k)$ and $n(k)$ are the clean signal and the additive noise in $x(k)$, respectively. Then denote a segment of noisy signal as $\{x(k)$, for $1 \leq k \leq L\}$ where $L = 1 + N_{ss}(K-1)$ is the total number of samples. Notation K is the number of samples used in GM(1,1) modeling and $N_{ss} = \lfloor L/(K-1) \rfloor$ is the number of subsets with one sample overlapped. The proposed grey filtering approach is given as follows.

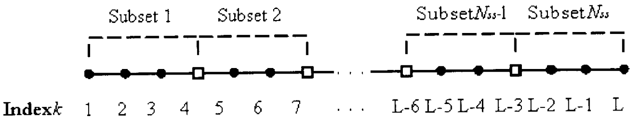


Fig. 1 One sample overlapped subsets for grey filtering.

Step 1: Divide $\{x(k)$, for $1 \le k \le L\}$ into N_{ss} subsets as $\{x_i(k)$, for $1 \le i \le N_{ss}\}$. The way to divide $x(k)$ into subsets for the case $K = 4$ is depicted in Fig. 1 where the square indicates the sample overlapped.

Step 2: For each subset i, find estimate of $x_i(k)$, $\hat{x}_i(k)$, by GM(1,1) model as stated in Sect. 2.1. Then consider $\hat{x}_i(k)$ as an estimate of $s_i(k)$, $\hat{s}_i(k)$. That is, $\hat{x}_i(k) = \hat{s}_i(k)$. Note that $\hat{x}_i(k) \ne \hat{s}_i(k)$ and therefore additive noise $n_i(k)$ is not equal to estimation error of GM(1,1) model, $e_i(k) = x_i(k) - \hat{x}_i(k) = x_i(k) - \hat{s}_i(k)$, in general.

Step 3: Since $e_i(k) \ne n_i(k)$ but related to $n_i(k)$, additive noise $n_i(k)$ is estimated as $\hat{n}_i(k) = \alpha e_i(k)$ where $\alpha > 0$ is a user-defined scaling parameter and is determined by experiences.

Step 4: Estimate mean μ of additive noise $n(k)$ as

$$\hat{\mu} = \frac{1}{N_{ss}(K-1)} \sum_{i=1}^{N_{ss}} \sum_{k=2+(i-1)(K-1)}^{1+i(K-1)} \hat{n}_i(k) \quad (15)$$

Since $x_i(k)$ is of one sample overlapped, thus only $\hat{n}(1) = 0$ is excluded in (15).

Step 5: Estimate standard deviation σ of $n(k)$ as

$$\hat{\sigma} = \left(\frac{1}{N_{ss}(K-1)} \times \sum_{i=1}^{N_{ss}} \sum_{k=2+(i-1)(K-1)}^{1+i(K-1)} (\hat{n}_i(k) - \hat{\mu})^2 \right)^{1/2} \quad (16)$$

To demonstrate that the proposed grey filtering approach is able to estimate additive noise through estimation error of GM(1,1) model, four examples are given in the following.

Example 1. constant signal without additive noise

The data sequence used here is $\{x(k)$, for $1 \le k \le 4\}$ and $x(k) = 5$ for all k. To fit in the additive signal model, $x(k)$ is expressed as $x(k) = x(k) + n(k)$ where $s(k) = 5$ and $n(k) = 0$. By GM(1,1) model, the estimated sequence $\hat{x}(k) = \hat{s}(k)$ is $\{5,\ 5,\ 5,\ 5\}$. Since $\hat{x}(1) = x(1)$, it is excluded in noise estimation. The estimated noise $\hat{n}(k)$ is obtained as $\{\hat{n}(2),\ \hat{n}(3),\ \hat{n}(3)\} = \{0,\ 0,\ 0\}$ where $\alpha = 1.0$. In this case, the scaling factor α has no effect on noise estimation. That is, it can be any appropriate positive number. This example indicates that $x(k)$ without additive noise can be estimated by GM(1,1) model precisely.

Example 2. random exponential signal without additive noise

A random signal is generated by the following equation:

$$x(k) = Ae^{-k/\tau} + n(k) \quad (17)$$

for $1 \le k \le 4$, where A and τ have uniform probability density functions (pdf) which varies from 2 to 3, and 10 to 20, respectively. The noise $n(k)$ is set to be zero on purpose. By (17), a sequence with four data samples is generated as $\{2.4372,\ 2.2786,\ 2.1303,\ 1.9916\}$. With $\alpha = 1.0$, the estimated noise $\hat{n}(k)$ through GM(1,1) model is obtained as $\{8.309 \times 10^{-4},\ 7.228 \times 10^{-4},\ 6.252 \times 10^{-4}\}$ which is approximately zero. Therefore, the random exponential signal can be estimated appropriately by GM(1,1) model.

Example 3. constant signal with additive noise

To illustrate that additive noise causes estimation error in GM(1,1) model, additive noise is put into account in Example 1. Here, the noise $n(k)$ is of Gaussian pdf with mean $\mu = 0$ and standard deviation $\sigma = 0.5$. By $x(k) = s(k) + n(k)$, 1,000 patterns are generated, where each pattern has 20 samples. By the proposed grey filtering approach with $K = 4$ and $\alpha = 1.0$, the estimated mean $\hat{\mu} = 0.0021$ and estimated standard deviation $\hat{\sigma} = 0.2914$ for the Gaussian additive noise $n(k)$. Obviously, the standard deviation σ of $n(k)$ is under estimated while mean μ is estimated appropriately. To estimate σ more accurate, after several trials the scaling factor $\alpha = 1.75$ is determined. Then 10 experiments are performed under the same conditions as given previously. The average of $\hat{\mu}$ and $\hat{\sigma}$ are 0.00376 and 0.50562, respectively. It is clear that statistics of $n(k)$ can be well estimated with the scaling factor $\alpha = 1.75$.

Example 4. random signal with additive noise

In this example, 1,000 patterns are generated by (17) for $1 \le k \le 20$, where A and τ have uniform pdf which varies from 2 to 3, and 10 to 20, respectively. The noise $n(k)$ is of Gaussian pdf with $\mu = 0$ and $\sigma = 0.1$. With $K = 4$ and $\alpha = 1.0$, the noise statistics are estimated as $\hat{\mu} = 7.4 \times 10^{-4}$ and $\hat{\sigma} = 0.0583$, respectively. Again, the standard deviation of $n(k)$ is under estimated. As before, the scaling factor α is set as 1.75 for better noise estimation. With $\alpha = 1.75$, 10 experiments are performed. The averages of $\hat{\mu}$ and $\hat{\sigma}$ are 0.0013 and 0.10138, respectively. Obviously, the estimation accuracy has been significantly improved.

3. Application of Grey Filtering to Speech Enhancement

In this section, the motivations for this paper are given in Sect. 3.1. Then the proposed spectral subtraction approach for speech enhancement, which is based on grey filtering, is described in Sect. 3.2.

3.1 Motivations

This paper is motivated by the following observations. As shown in Sect. 2.2, the estimation error of GM(1,1) model is zero or approximately zero when additive noise is absent and non-zero when additive noise is included. This is true both for constant and random exponential signal. This implies that the estimation error of GM(1,1) model can be related to additive noise. As demonstrated in Sect. 2.2, statistics of additive noise can be estimated accurately with an appropriate scaling factor α. Next, a clean speech signal generally consists of non-speech and speech portions. The non-speech portion can be considered as constant signal while the speech portion as random signal. Consequently, there is a hope that the estimation error of GM(1,1) model for clean speech is approximate to zero and non-zero when noisy speech is present. Moreover, there is no need to determine speech and non-speech portion as in [2] since both constant and random signals can be estimated appropriately and the GM(1,1) modeling requires as few as four data samples. To demonstrate the idea just described, the clean speech b.wav (male speech of letter "b") obtained from [8] is provided as an example which is shown in Fig. 2 (a). Since b.wav is within the range $(-1, 1)$, it fails to meet the requirement of Condition (i) in Sect. 2.1. To make it satisfied, b.wav is level-shifted by 5 before it is put into GM(1,1) modeling. Condition (ii) is met in speech signal since adjacent samples does not change abruptly in general. The estimate of b.wav obtained from GM(1,1) model is given in Fig. 2 (b) where $K = 4$. Obviously, the estimate of b.wav by GM(1,1) model retains b.wav appropriately.

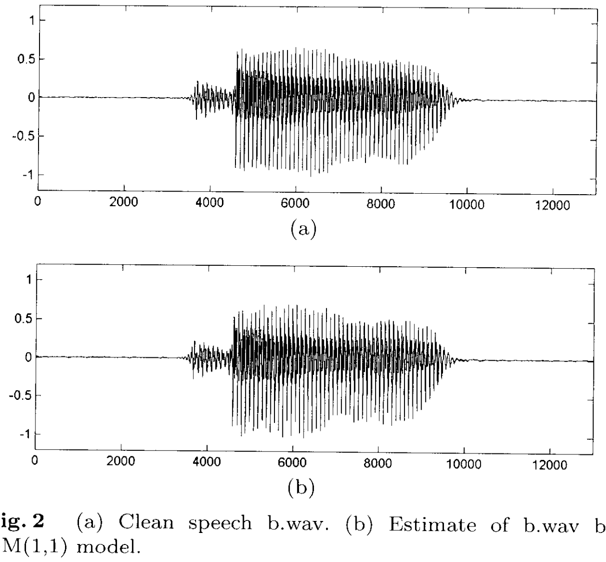


Fig. 2 (a) Clean speech b.wav. (b) Estimate of b.wav by GM(1,1) model.

Note that the standard deviation σ of additive noise $n(k)$ can be estimated accurately by grey filtering and that a spectral subtraction approach for speech enhancement depends heavily on the accuracy of the standard deviation of $n(k)$. An MSS [7] approach based on grey filtering is proposed in this paper. The proposed approach is described in the following subsection.

3.2 The MSS Approach Based on Grey Filtering

Assume that the additive signal model is appropriate for the noisy speech and that the noisy speech signal is stored in the wave file format whose range is within $(-1, 1)$. The diagram block for the proposed magnitude spectral subtraction (MSS) approach based on grey filtering is depicted in Fig. 3. Given a noisy speech signal $x_o(k) = s_o(k) + n_o(k)$, the implementation steps are described in the following where additive noise $n_o(k)$ is assumed known and the length of $x_o(k)$ is assumed as a multiple of L.

Step 1: Shift up the level of $x_o(k)$ by an appropriate constant C, $x_o(k) \leftarrow x_o(k) + C$, such that Condition (i) is met.

Step 2: Divide $x_o(k)$ into M non-overlapped segments of length L and denote $x(k)$ as speech segment of length L. Then Steps 3 to 9 are performed for each speech segment $x(k)$.

Step 3: With rectangular window, obtain $X(f) = FFT_L\{x(k)\} = S(f) + N(f)$ where $FFT_L\{\cdot\}$ denotes as L-point fast Fourier transform (FFT).

Step 4: Estimate additive noise $n(k)$ as $\hat{n}(k)$ by the grey filtering approach described in Sect. 2.2.

Step 5: Perform L-point FFT on $\hat{n}(k)$ to find the magnitude of $\hat{N}(f)$, $|\hat{N}(f)|$, where $\hat{n}(1) = \hat{n}(2)$ is used.

Step 6: Estimate the standard deviation of $|\hat{N}(f)|$, $\sigma_{|\hat{N}(f)|}$.

Step 7: Perform MSS [7] as

$$|\hat{S}(f)| = \begin{cases} D = |X(f)| - \beta\sigma_{|\hat{N}(f)|}, & \text{if } D > 0 \\ 0, & \text{else} \end{cases} \tag{18}$$

where $|\hat{S}(f)|$ is an estimate of $|S(f)|$ and β is a user-defined scaling factor determined by experiences.

Step 8: Find estimate of $s(k)$ as $\hat{s}(k) = IFFT_L\{|\hat{S}(f)| e^{j\angle X(f)}\}$ where $IFFT_L\{\cdot\}$ is the inverse of L-point FFT and $\angle X(f)$ is the angle obtained by performing $FFT_L\{\cdot\}$ on noisy speech segment $x(k)$.

Step 9: Shift down the level of $\hat{s}(k)$ by the constant C.

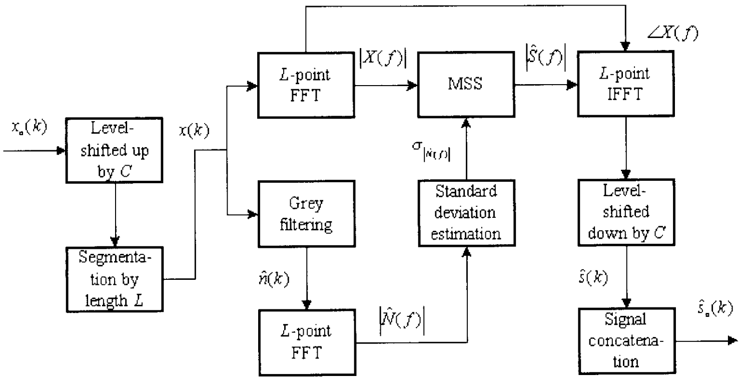


Fig. 3 The block diagram for the proposed MSS approach.

Step 10: Concatenate M segments $\hat{s}(k)$ to find estimate of $s_o(k)$, $\hat{s}_o(k)$.

Step 11: Obtain residual noise $n_r(k) = \hat{s}_o(k) - s_o(k)$.

Step 12: Calculate input, output, and improvement signal to noise ratios, SNR_{in}, SNR_{out}, and SNR_{imp}, as follows:

$$SNR_{in} = 10\log\frac{\sum_{k=1}^{M\times L} s_o^2(k)}{\sum_{k=1}^{M\times L} n_o^2(k)} \tag{19}$$

$$SNR_{out} = 10\log\frac{\sum_{k=1}^{M\times L} \hat{s}_o^2(k)}{\sum_{k=1}^{M\times L} n_r^2(k)} \tag{20}$$

$$SNR_{imp} = SNR_{out} - SNR_{in} \tag{21}$$

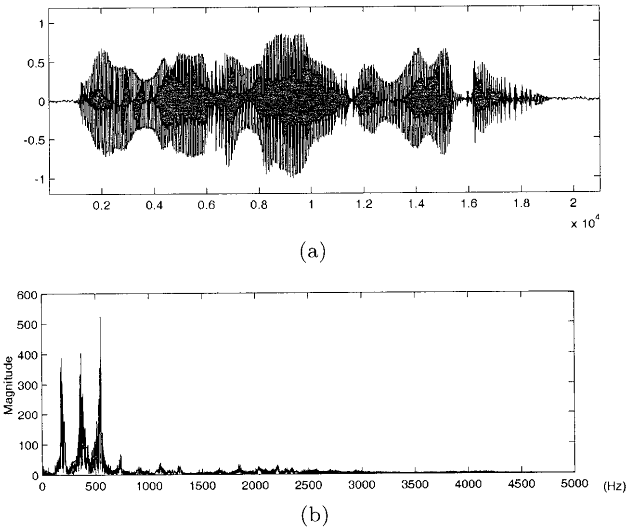
There are at least three advantages in the proposed MSS approach. First, the mechanism to determine non-speech and speech portions as in [2] is not required. Second, there is no need to trade bandwidth and memory capacity, such as higher sampling rate in [4] and zero insertion in [5], for noise estimation. Third, in the proposed MSS approach no re-sampling scheme is needed as in [4] and no synchronization is required as in [5].

4. Simulation Results, Discussions, and Comparisons

In this section, the proposed MSS approach for speech enhancement described in Sect. 3.2 is justified. Two examples are provided in Sect. 4.1 to investigate the performance of the proposed approach. Then discussions on simulation results are given in Sect. 4.2. Finally, the proposed MSS approach is compared with high frequency region based (HFR-based) approach in [4] and zero-padding (ZP) approach in [5] in Sect. 4.3.

4.1 Simulation Results

By using MATLAB, the proposed MSS approach depicted in Fig. 3 is programmed. Two examples are given in the simulation. In the first example, speech file f0125s.wav in [8] is used which is a female oral reading the sentence "We were away a year ago." For more details, one may consult in the Appendix 4 of [8]. The sampling rate for f0125s.wav is 10 KHz and the length of samples is 21,000. In the simulation, the speech file f0125s.wav is level-shifted by 5, i.e., $C = 5$ and the segment length is set to 1,000. That is, $L = 1{,}000$ and therefore the number of segments $M = 21{,}000/1{,}000 = 21$. The number of samples used in GM(1,1) modeling is 4, i.e., $K = 4$. And the scaling factor $\beta = 5$ in (18) is used. In the additive signal model $x_o(k) = s_o(k) + n_o(k)$, the file f0125s.wav shown in Fig. 4 (a) is considered as clean speech $s_o(k)$ whose spectrum is given in Fig. 4 (b), and the additive noise $n_o(k)$ is artificially generated. Two types of additive noise are generated. One is Gaussian noise and the other is uniform noise. The Gaussian noise is generated with zero mean and a specified standard deviation σ. With different σ, the SNR_{in}, SNR_{out}, and SNR_{imp} for each case are given in Fig. 5. The noisy speech and enhanced speech for the Gaussian case with $\sigma = 0.05$ are shown in Figs. 6 (a) and 6 (b), respectively. The corresponding spectra for noisy speech and enhanced



(a)

(b)

Fig. 4 (a) Clean speech. (b) Clean spectrum (f0125s.wav).

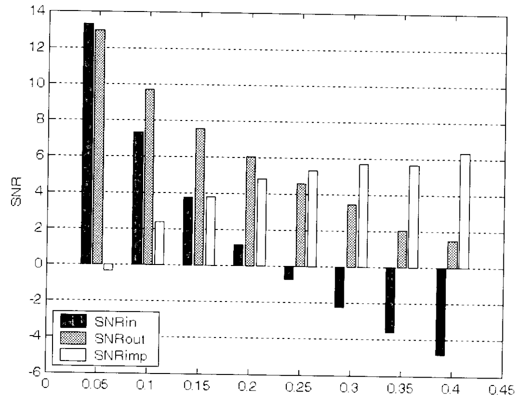
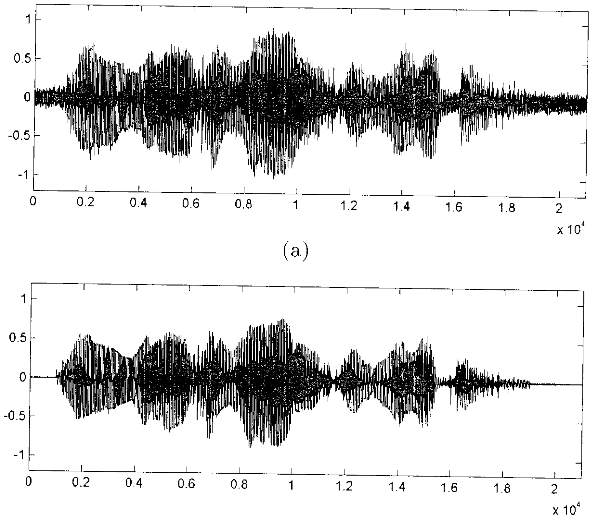


Fig. 5 SNR_{in}, SNR_{out}, and SNR_{imp}, for Gaussian noise with different σ (f0125s.wav).



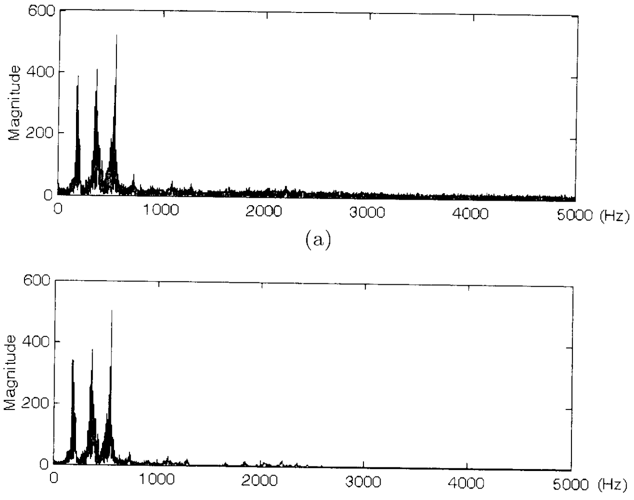
(a)

(b)

Fig. 6 (a) Noisy speech (σ = 0.05). (b) Enhanced speech (f0125s.wav).

speech are given in Figs. 7 (a) and 7 (b), respectively. As a second type of additive noise, the uniform noise is distributed within the range $\gamma(-0.5,\ 0.5)$ where γ is a scaling factor. The SNR_{in}, SNR_{out}, and SNR_{imp}, for several values of γ are depicted in Fig. 8. The noisy speech $x_o(k)$ and the enhanced speech $\hat{s}_o(k)$ for $\gamma = 0.4$ are, respectively, shown in Figs. 9 (a) and 9 (b) whose corresponding spectra are given in Fig. 10.

The second example used in the simulation is the speech file f0101s.wav in [8] which is a female speech counting from one to ten. The clean speech and spectrum of f0101s.wav are shown in Fig. 11. The sampling rate is 10 KHz and the length of samples is 98,000. In the simulation, parameters $C = 5$, $L = 1{,}000$, $M = 98$, $\beta = 5$, and $K = 4$. The speech and its spectrum contaminated by Gaussian noise with $\sigma = 0.1$ and the enhanced speech are given in Fig. 12. The spectra corresponding to Fig. 12 are showed in Fig. 13. The



(a)

(b)

Fig. 7 (a) Noisy spectrum (σ = 0.05). (b) Enhanced spectrum (f0125s.wav).

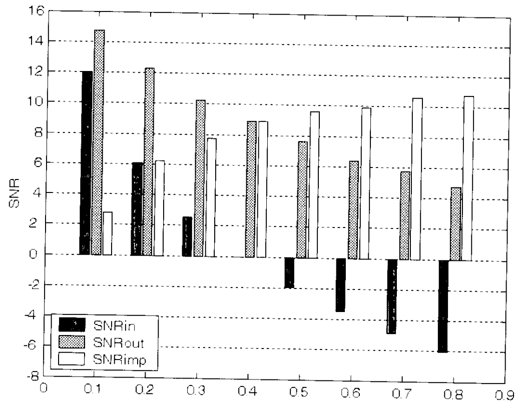
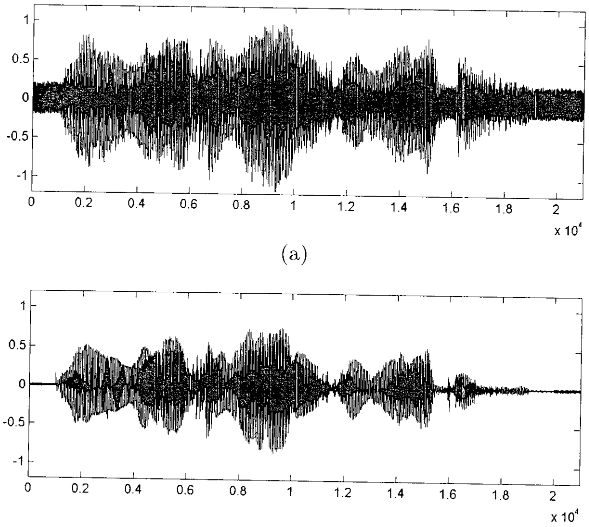


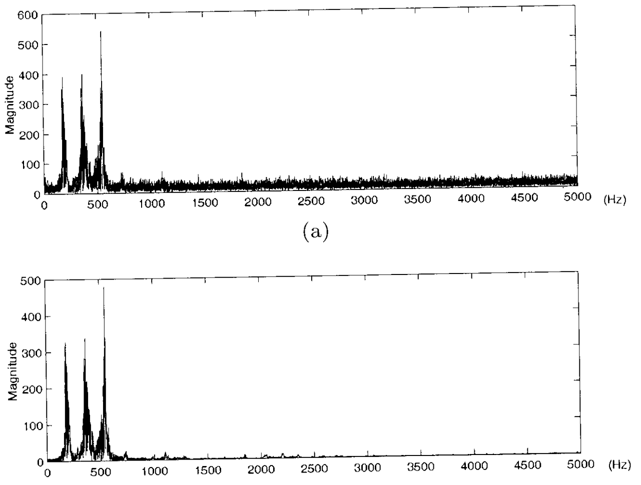
Fig. 8 SNR_{in}, SNR_{out}, and SNR_{imp}, for uniform noise with different γ (f0125s.wav).



(a)

(b)

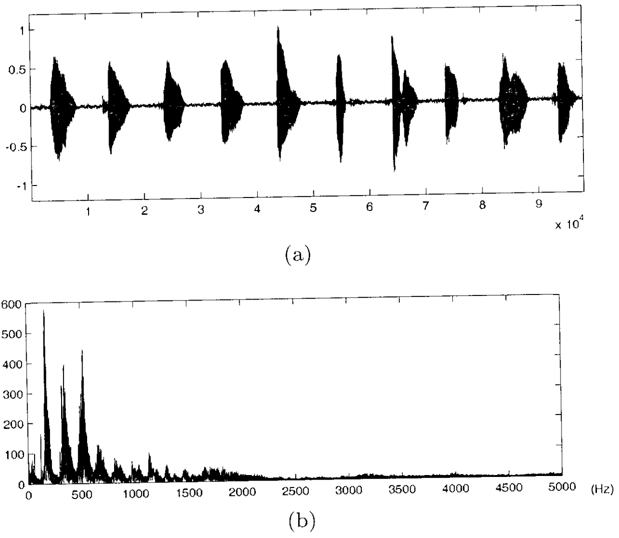
Fig. 9 (a) Noisy speech (γ = 0.4). (b) Enhanced speech (f0125s.wav).



(a)

(b)

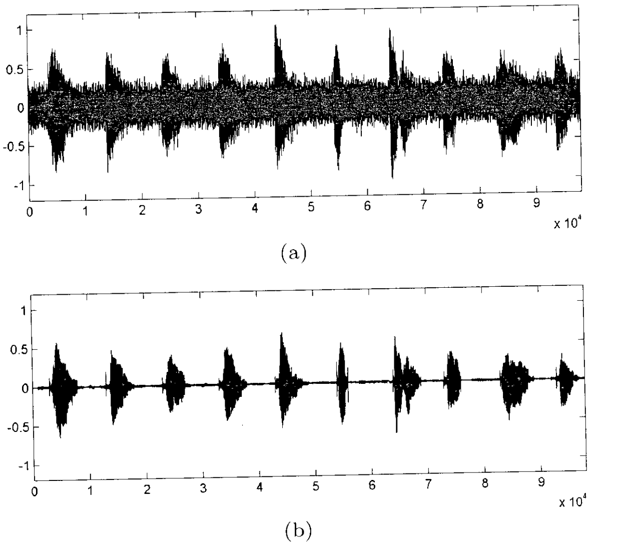
Fig. 10 (a) Noisy spectrum ($\gamma = 0.4$). (b) Enhanced spectrum (f0125s.wav).



(a)

(b)

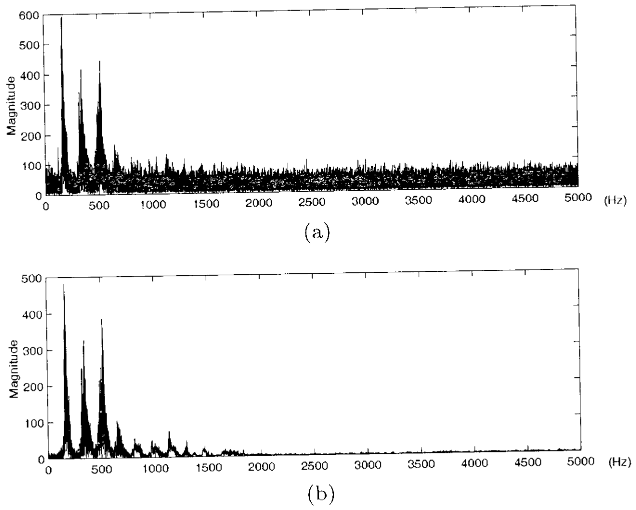
Fig. 11 (a) Clean speech. (b) Clean spectrum (f0101s.wav).



(a)

(b)

Fig. 12 (a) Noisy speech ($\sigma = 0.1$). (b) Enhanced speech (f0101s.wav).



(a)

(b)

Fig. 13 (a) Noisy spectrum ($\sigma = 0.1$). (b) Enhanced spectrum (f0101s.wav).

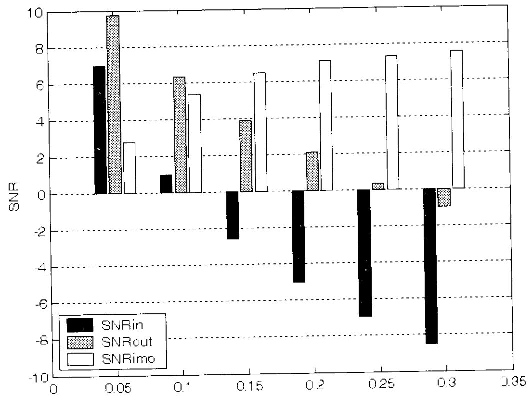


Fig. 14 SNR_{in}, SNR_{out}, and SNR_{imp}, for Gaussian noise with different σ (f0101s.wav).

SNR_{in}, SNR_{out}, and SNR_{imp} for different σ are plotted in Fig. 14. Figure 15 summarizes the SNR_{in}, SNR_{out}, and SNR_{imp} for uniform noise with different γ. The noisy speech and enhanced speech, with their corresponding spectra, for the case of $\gamma = 0.4$ are depicted in Figs. 16 and 17, respectively.

4.2 Discussions

The simulation results, shown in Figs. 5, 8, 14, and 15, indicate that the proposed MSS approach improves *SNR* in different degrees except the case with Gaussian noise of $\sigma = 0.05$ in Fig. 5. To investigate the case, energies of $s_o(k)$, $n_o(k)$, $\hat{s}_o(k)$, and $n_r(k)$ are calculated and their values are 1,121.40, 52.60, 923.99, and 47.00, respectively. The energy loss for the speech component is $1{,}121.40 - 923.99 = 197.41$ while the energy of noise is suppressed by $52.60 - 47.00 = 5.60$. This implies that the proposed approach to remove the portion of energy

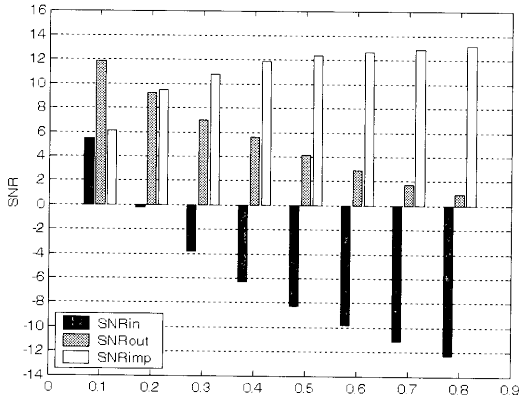
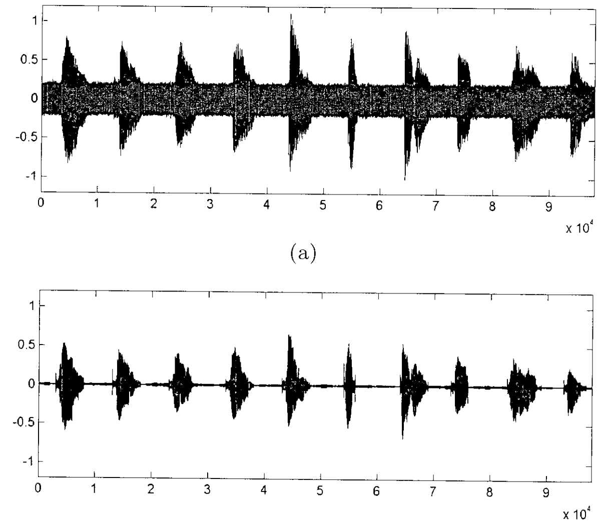


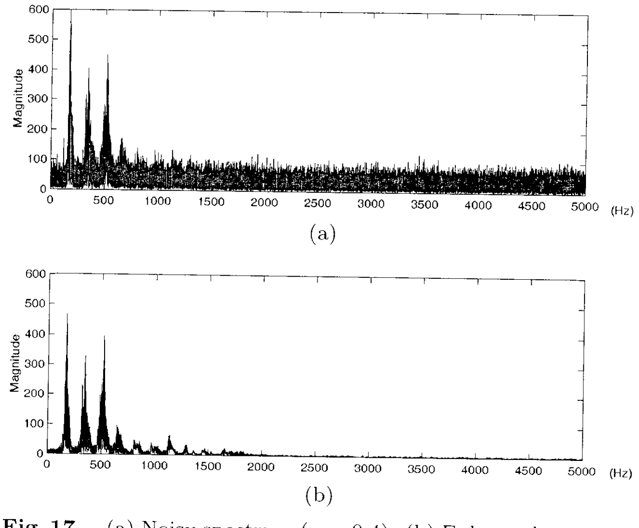
Fig. 15 SNR_{in}, SNR_{out}, and SNR_{imp}, for uniform noise with different γ (f0101s.wav).



(a)

(b)

Fig. 16 (a) Noisy speech (γ = 0.4). (b) Enhanced speech (f0101s.wav).



(a)

(b)

Fig. 17 (a) Noisy spectrum ($\gamma = 0.4$). (b) Enhanced spectrum (f0101s.wav).

in $s_o(k)$ is more than that in $n_o(k)$ for the case. Therefore, SNR_{in} is higher than SNR_{out}. Though SNR_{imp} is negative, the speech shown in Fig. 6 (b) has better speech quality. That is, the noisy speech in Fig. 6 (a) is enhanced and thus the objective of speech enhancement is achieved by the proposed MSS approach even it is of little degradation in *SNR*.

The second observation from simulation results in Figs. 5, 8, 14, and 15 is that the proposed MSS approach has better results for uniform noise than that for Gaussian noise. In other words, for similar SNR_{in}, SNR_{imp} for uniform cases are higher than that for Gaussian cases. This is true both for files f0125s.wav. and f0101s.wav. For example, compare the case $\sigma = 0.35$ in Fig. 5 with the case $\gamma = 0.6$ in Fig. 8. The SNR_{in} are −3.60 dB and −3.50 dB, respectively. However, the SNR_{imp} for the case $\gamma = 0.6$ is higher than the case $\sigma = 0.35$ by 4.24 dB. This suggests that the proposed MSS approach is favorable to uniform additive noise.

Finally, note that the SNR_{imp} for the Gaussian case of $\sigma = 0.05$ in Fig. 14 is positive instead of negative as in Fig. 5. One possible reason for this is that the major portion of f0125s.wav is speech while the portions of non-speech and speech in f0101s.wav are approximately equal. It implies that the energy in f0101s.wav is less than that in f0125s.wav. In other words, for a given amount of additive noise SNR_{in} in f0101s.wav is less than that in f0125s.wav. This can be verified from Figs. 5 and 8, or Figs. 14 and 15. Consequently, the portion of energy loss resulted from the proposed MSS approach in f0101s.wav is less than that in f0125s.wav in general. This may explain why the SNR_{imp} is positive for the Gaussian case of $\sigma = 0.05$ in Fig. 14.

4.3 Comparison with HFR-Based and ZP Approaches

In this subsection, the performance of proposed MSS approach is compared with HFR-based approach in [4] and ZP approach in [5] in terms of SNR_{imp} and subjective listening quality, respectively.

4.3.1 Objective Comparison Results

The simulation for HFR-based approach is described here. First, both speech files f0125s.wav and f0101s.wav are re-sampled where the new sampling rate is 30 KHz. Then additive noise is generated and re-sampled with sampling rate 30 KHz. Next, the following steps are implemented: (i) Divide noisy speech with new sampling rate into non-overlapped segments of length $L = 1{,}000$ as in the proposed MSS approach. (ii) For each noisy speech segment perform L-point FFT with rectangular window. (iii) Calculate the mean of magnitudes in the range from 10 KHz to 15 KHz. (iv) Perform MSS by subtracting the mean obtained in Step (iii) from the magnitude found in Step (ii). (v) Perform L-point IFFT, with the estimated angles obtained in Step (ii),

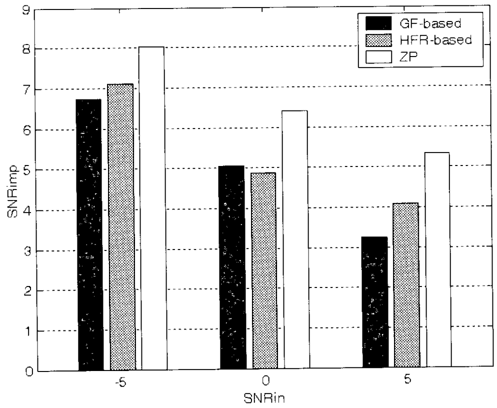


Fig. 18 Comparison results of SNR_{imp} for Gaussian noise (f0125s.wav).

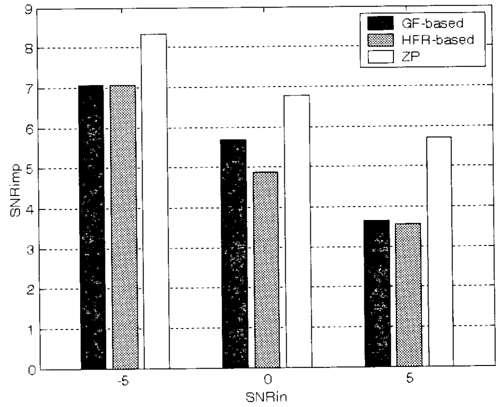


Fig. 20 Comparison results of SNR_{imp} for Gaussian noise (f0101s.wav).

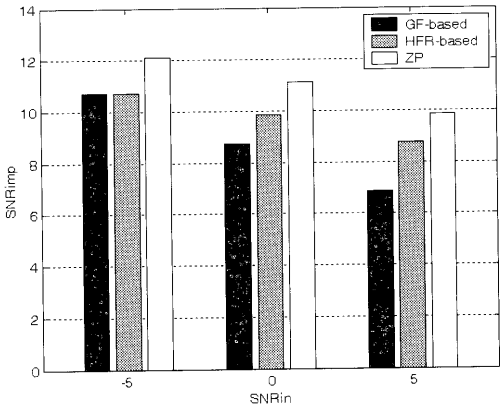


Fig. 19 Comparison results of SNR_{imp} for uniform noise (f0125s.wav).

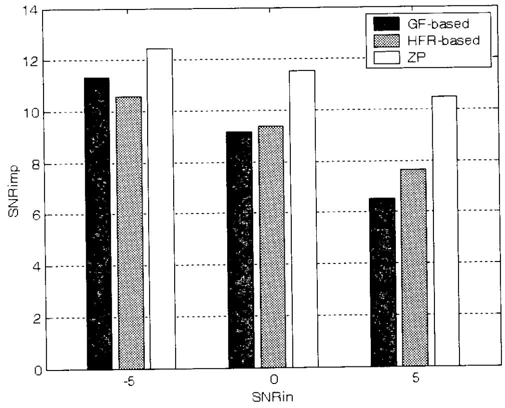


Fig. 21 Comparison results of SNR_{imp} for uniform noise (f0101s.wav).

to reconstruct speech segment. (vi) Concatenate reconstructed speech segments which is then down sampled by a factor of 3. (vii) Calculate SNR_{imp} as in (21) where the speech and additive noise before re-sampling are used to find SNR_{in}. By the steps just described, three cases of SNR_{in}, -5 dB, 0 dB, and 5 dB, are considered in the simulation. The resulted SNR_{imp} of speech files f0125s.wav and f0101s.wav, for Gaussian and uniform additive noises, are shown in Figs. 18, 19 and Figs. 20, 21, respectively.

As for the ZP approach, the simulation is performed as follows: First, zero samples are inserted between samples in f0125s.wav and f0101s.wav, respectively. Then additive noise is generated and added to zero-padded speech to form noisy speech, which is of twice length of the original speech. Next, the following steps are performed: (i) Divide noise speech into segments of length $L = 1{,}000$ as in the proposed MSS approach. (ii) Estimate additive noise by the samples where zero samples are inserted. (iii) For each noisy speech segment perform L-point FFT with rectangular window. (iv) Transform the estimated noise by L-point FFT and calculate the mean of magnitudes. (v) Perform MSS by subtracting the mean obtained in Step (iv) from the magnitude found in Step (iii). (iv) Perform L-point IFFT, with the estimated angles obtained in Step (iii), to reconstruct speech segment. (vi) Concatenate reconstructed speech segments and discard the samples where zero padding is operated. (vii) Calculate SNR_{imp} as in (21) where the original speech and additive noise added to the original speech are used to find SNR_{in}. By setting SNR_{in} to -5 dB, 0 dB, and 5 dB, the simulation results of SNR_{imp} for speech files f0125s.wav and f0101s.wav, both for Gaussian and uniform additive noises cases, are depicted in Figs. 18, 19 and Figs. 20, 21, respectively.

The comparison results for the proposed MSS approach, HFR-based approach, and ZP approach are shown from Fig. 18 to Fig. 21 where the proposed approach is denoted as GF-based approach. On average, the proposed approach is inferior to HFR-based approach in SNR_{imp} by 0.6770 dB for file f0125s.wav while

0.0557 dB superior for file f0101s.wav. When compared with ZP approach, the proposed approach is inferior, on average, by 1.9208 dB in SNR_{imp} for file f0125s.wav and 1.9719 dB inferior for file f0101s.wav. To sum up, the proposed MSS approach is slightly better than the HFR-based approach for file f0101s.wav and a little bit worse for file f0125s.wav in terms of SNR_{imp}. As compared with ZP approach, the proposed approach is inferior by approximately 1.95 dB in SNR_{imp} for both files. However, it should be noted that both HFR-based and ZP approaches require increase in transmission bandwidth. Twice the original bandwidth is demanded in ZP approach and three times in HFR-based in the simulation. On the other hand, the proposed MSS approach use available noisy speech as it is without re-sampling or zero insertion. Therefore, to some extend it can be said that HFR-based and ZP approaches trade bandwidth for better objective performance. Besides, the receiver requires to be synchronized with the transmitter for better performance in ZP approach. Consequently, it is an appropriate choice to use the proposed MSS approach instead of HFR-based or ZP approaches when bandwidth increase is not possible in the application of interest.

4.3.2 Subjective Comparison Results

It is known that better SNR, an objective assessment, does not mean better listening quality in general. Thus, in this subsection, the subjective listening quality is compared for the proposed approach, HFR-based approach, and ZP approach. As in [4], the quality of test speech is divided into five levels: (i) very good, (ii) good, (iii) normal, (iv) bad, and (v) very bad. The five levels are then scored as follows: 5 to level (i), 4 to level (ii), 3 to level (iii), 2 to level (iv), and 1 to level (v). For files f0125s.wav and f0101s.wav, three cases of SNR_{in}, -5 dB, 0 dB, and 5 dB, are considered in the listening test. Twelve persons with normal hearing ability are involved to give the score for each approach where additive noise is Gaussian or uniform. The scores for different cases are then averaged, respectively. The means of scores for different SNR_{in} and additive noise are shown in Figs. 22 and 23 for file f0125s.wav and Figs. 24 and 25 for file f0101s.wav. Interesting enough, all test results indicate that the proposed MSS approach has better subjective listening quality even though its SNR_{imp} is worse than HFR-based and ZP approaches in most of cases as shown in the previous subsection. This interesting issue can be investigated through Eq. (20). In (20), the residual noise $n_r(k)$ attempts to indicate the residual of additive noise. However, it is not an appropriate indication to the residual of additive noise when an over-subtraction happens in MSS. For example, assume the maximum magnitude in additive noise is λ. Then a value greater than λ is used in MSS to subtract from the magnitude of noisy spectrum. It is obvious that all additive noise is removed but $n_r(k) \neq 0$. In this case, $n_r(k)$ is related to the signal loss instead of additive noise since all additive noise has been removed. In general, an over-subtraction in MSS results in musical noise [7], which is a tin-like sound, in the enhanced

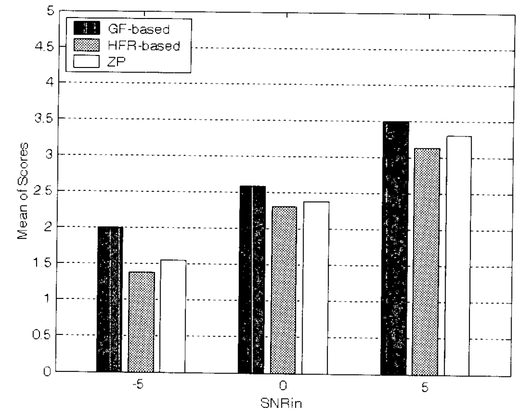


Fig. 22 Listening test results for Gaussian noise (f0125s.wav).

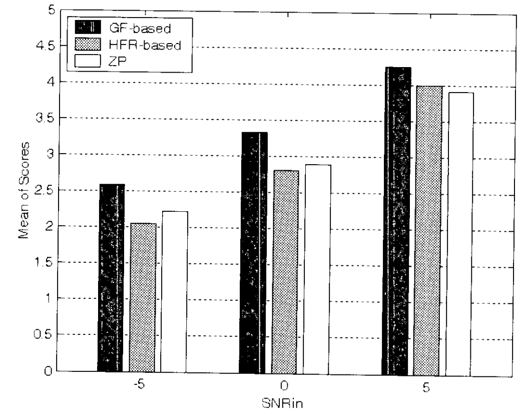


Fig. 23 Listening test results for uniform noise (f0125s.wav).

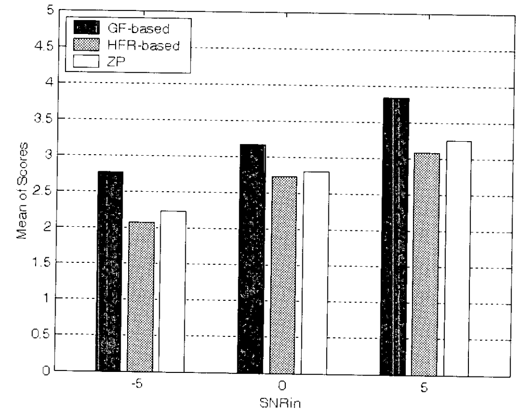


Fig. 24 Listening test results for Gaussian noise (f0101s.wav).

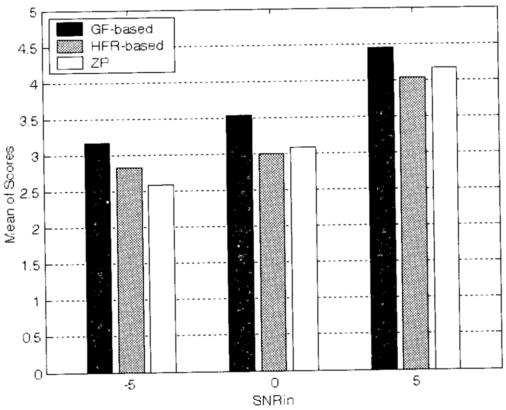


Fig. 25 Listening test results for uniform noise (f0101s.wav).

speech. On the other hand, $n_r(k)$ is generally related to the residual of additive noise when an over-subtraction is not the case in MSS. In this case, $n_r(k)$ results in annoying sound of additive noise in the enhanced speech in general. To justify the idea described above, the case of $SNR_{in} = 0\,\text{dB}$ in f0101s.wav with uniform additive noise is given as an example. First, energies of $s_o(k)$ and $n_o(k)$ are calculated and their values are 1,227.30 and 1,242.10, respectively. Then the energies of $\hat{s}_o(k)$ and $n_r(k)$ in the proposed MSS approach, HFR-based approach, and ZP approach are found separately. Energies of $\hat{s_o}(k)$ and $n_r(k)$ are 880.18 and 106.26 in the proposed approach, 1,130.00 and 132.05 in HFR-based approach, and 1,110.30 and 78.72 in ZP approach, respectively. For the proposed MSS approach, the energy of $\hat{s}_o(k)$ is far less than that in $s_o(k)$. It implies that an over-subtraction is taken place quite possibly and therefore $n_r(k)$ is related to the signal loss instead of additive noise in general. The over-subtraction is verified in the listening test where musical noise is heard as expected. As for HFR-based and ZP approaches, the energy of $\hat{s}_o(k)$ is close to $s_o(k)$. Consequently, $n_r(k)$ reflects the residual of additive noise since an over-subtraction in MSS may not happen and an annoying sound of additive noise could be heard in the listening test. As expected, the enhanced speech obtained in HFR-based or ZP approach reveals an annoying sound of additive noise in the listening test. For other cases in Figs. 18 to 21, we have similar results. That is, in the proposed MSS approach a tin-like sound of musical noise can be heard in the enhanced speech and an annoying sound of additive noise heard both in HFR-based and ZP approaches. Since people have less complaint on musical noise when compared with additive noise, thus the proposed MSS approach has better subjective listening quality than HFR-based and ZP approaches. Besides, the listening test results suggest that all three approaches are favorable to uniform additive noise. This can be justified in Figs. 18 to 21.

5. Conclusive Remarks

In this paper, a grey filtering approach based on GM(1,1) model is proposed. Then the proposed grey filtering approach is applied to speech enhancement whose noise removal technique is MSS. This paper is motivated by the following observations. For constant signal and random signal, GM(1,1) model has zero or approximately zero estimation error when additive noise is absent and non-zero when additive noise is present. Therefore, the estimation error of GM(1,1) model is related to additive noise in the proposed grey filtering approach. This idea is verified by several examples. The simulation results show that the proposed grey filtering is able to estimate additive noise appropriately. Next, note that the speech signal generally consists of non-speech and speech portions. The non-speech portion can be considered as constant signal while speech portion as random signal. Thus, an MSS-based speech enhancement approach based on grey filtering is proposed. Then the proposed MSS approach is justified by two examples where Gaussian noise and uniform noise are considered. The simulation results indicate that the proposed MSS approach works well for both cases and is favorable to case of uniform noise. Besides, the proposed MSS approach is compared with HFR-based approach in [4] and ZP approach in [5] in terms of SNR_{imp} and subjective listening quality. The simulation results show that the proposed MSS approach has worse performance in SNR_{imp} than HFR-based and ZP approaches in most of cases. However, the proposed MSS approach has better subjective listening quality than HFR-based and ZP approaches.

Acknowledgement

This work was supported by the National Science Council of the Republic of China under grants NSC 90-2213-E-270-006 and NSC 91-2213-E-270-002.

References

[1] S. Haykin, Adaptive Filtering Theory, 3rd ed., Prentice-Hall, 1996.

[2] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE, vol.67, no.12, pp.1689–1697, Nov. 1962.

[3] W. Kim and H. Ko, "Noise variance estimation for Kalman filtering of noisy speech," IEICE Trans. Inf. & Syst., vol.E84-D, no.1, pp.155–160, Jan. 2001.

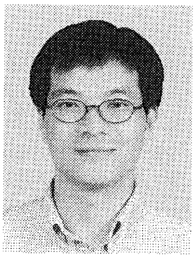
[4] J. Yamauchi and T. Shimamura, "Noise estimation using high frequency regions for spectral subtraction," IEICE Trans. Fundamentals, vol.E85-A, no.3, pp.723–727, March 2002.

[5] L. Singh and S. Sridharan, "Speech enhancement using preprocessing," Proc. IEEE TENCON, vol.2, pp.755–758, 1997.

[6] J. Deng, "Introduction to grey system theory," J. Grey System, vol.1, pp.1–24, 1989.

[7] S.V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, 2nd ed., John Wiley & Sons, 2000.

[8] D.G. Childers, Speech Processing and Synthesis Toolboxes, John Wiley & Sons, 1999.



Cheng-Hsiung Hsieh was born in Nantou, Taiwan, in 1963. He received the B.S. degree in Electronic Engineering from National Taiwan Institute of Technology, Taiwan, in 1989. In 1995, he earned the M.S. degree from the Department of Electrical Engineering of Tennessee Technological University, USA. He obtained his Ph.D. degree in Electrical Engineering from the University of Texas at Arlington, USA, in 1997. Currently, he is a associate professor at Department of Electronic Engineering in Chien Kuo Institute of Technology, Changhua, Taiwan, Republic of China. His research interests include grey systems, artificial neural networks, digital image processing, statistical signal processing, and digital communications.