A New Soft Masking Method for Speech Enhancement in the Frequency Domain

Huan Zhao[1], Jun Liu[1], Zuo Chen[1], Fei Wang[1]
[1]*School of Information Science and Engineering, Hunan University,*
Lushan South Rood, Changsha 410082, P. R. China
liujun543@126.com

***Abstract*—Recently, ideal binary mask (IdBM) method has attracted keen interest because of its superiority in improving speech intelligibility. This method processes noisy speech based on time-frequency (T-F) unit. If the local Signal to Noise Ratio (SNR) is higher than the threshold, the T-F unit is retained; else, the T-F unit would be removed. This method works well in computational auditory scene analysis (CASA) field. However, as the threshold is usually low, much residual noise would exist. In addition, the accurate local SNR is difficult to obtain in practice. In this paper, we try to propose a new method to improve speech quality and intelligibility. Instead of finding a new way to estimate the local SNR, we try to compute the probability of local SNR higher than the threshold. After that, we multiply T-F units with a proper value to compress the residual noise. Results from sufficient experiments showed that our method performs well.**

***Index Terms*—Ideal binary mask (IdBM), threshold, speech quality and intelligibility, residual noise.**

I. INTRODUCTION

As speech signal can be easily polluted by kinds of noises, speech enhancement has become an important means of speech signal processing. For application like Automatic Speech Recognition (ASR) system, whether speech enhancement is adopted can make a big difference.

Many classic and effective speech enhancement methods have been proposed in the past. These algorithms such as Wiener filtering [1] and minimum mean-square error (MMSE) [2] can greatly improve the speech quality, but in terms of speech intelligibility, few gains were obtained. In [3], Loizou revealed that the amplification distortions exceeding 6 dB should be responsible for damage of speech intelligibility. Meanwhile, most of the existing algorithms allow this kind of amplification distortion. Recently, many people begin to focus on the IdBM method which is usually used in the CASA [4]. After using it to process noisy speech, the speech intelligibility can be improved markedly according to [5], [6]. The realization of IdBM method can be regarded as binary masking. Speech signals are first transformed to the frequency domain and divided into many T-F units. Then, a proper threshold would be selected. When the corresponding local SNR is higher than the selected threshold, the T-F unit should be saved, otherwise, the T-F unit should be abandoned, which means the gain value is zero. We notice that the threshold is usually low, which means much noise would be reserved. This certainly is harmful to speech quality and intelligibility. In this paper, we try to modify the binary gain function to compress noise. More specifically, we choose new gain value form to replace one when local SNR is higher than the threshold.

There is still a problem needs to consider. As we can see form definition, the IdBM method needs accurate local SNR to decide gain value, which is very difficult in practice. In [4], the author proposed many advanced and useful algorithms, of particular interest is the SMPO method. Instead of trying to calculate accurate local SNR, this method calculates the probability of local SNR greater than the threshold.

The organizational structure of this paper is as follows. Section II introduces the background knowledge and assumptions. Section III describes the details of the proposed method. In Section IV, we present the experimental details and the experimental results. The conclusion is drawn in Section V.

II. ASSUMPTIONS AND HYPOTHESIS MODEL

For speech enhancement, the first step is to establish a proper analysis model. Many research used the linear additive model. Based on this model, degraded speech signal is the sum of the clean speech signal and noise signal

$$y(t) = x(t) + n(t), \tag{1}$$

where $x(t)$ and $n(t)$ represent clean speech signal and noise signal, $y(t)$ is the degraded speech signal. The above model is the most common model in the speech signal processing area, but methods based on this model usually have high computational complexity. Apart from the linear additive model, there is another basic model which is widely used in spectral subtraction algorithms, according to this model, the power spectrum of the clean speech signal $P_x(\omega)$ plus the power spectrum of noise signal $P_n(\omega)$ is equal to the power spectrum of degraded speech signal $P_y(\omega)$

$$P_y(\mu) = P_x(\mu) + P_n(\mu), \tag{2}$$

We know this model is reasonable only if $x(t)$ and $n(t)$ are

Manuscript received May 29, 2013; accepted September 24, 2013.

This work was sponsored by National Nature Science Foundation of China (61173106), Specialized Research Fund for the Doctoral Program of Higher Education, China (20100161120021) and the Young Teacher's Growth Program of Hunan University.

uncorrelated stationary random processes. Even so, it can deduce some useful methods.

Applying the short-time Fourier transform on $x(t)$, $n(t)$ and $y(t)$, we get frequency domain signals, i.e., $X(k,\tau)$, $N(k, \tau)$, and $Y(k, \tau)$. Then, we use the magnitude-squared spectrum to approximate the power spectrum. This approximation is common in spectral subtraction algorithms [7]–[12]. Then, (2) can be rewritten as

$$Y\left(k,{}^{2}\right)^{2} \approx X\left(k,{}^{2}\right)^{2} + N\left(k,{}^{2}\right)^{2}. \tag{3}$$

Equation (3) is simplified as follows for convenience

$$Y_k^2 \approx X_k^2 + N_k^2. \tag{4}$$

As we all known, the short-time Fourier transform coefficients can be divided into two parts, i.e., the real part and imaginary part. Here we assume both the real part and the imaginary part obey Gaussian distribution and have equal variance, besides, the two distributions are independent [13], [14]. In this condition, the probability densities of $X_k^2$and N_k^2 obey exponential distribution according to [15], the corresponding probability distribution functions (PDF) are given by:

$$f_{X_k^2}\left(X_k^2\right) = \left[1/\pm_x^2(k)\right]\cdot\exp\left[-X_k^2/\pm_x^2(k)\right], \tag{5}$$

$$f_{N_k^2}\left(N_k^2\right) = \left[1/\pm_n^2(k)\right]\cdot\exp\left[-N_k^2/\pm_n^2(k)\right], \tag{6}$$

where $\sigma_x^2(k)$ and $\sigma_n^2(k)$ denote clean speech and noise variance, respectively. According to Characteristics of exponential distribution, the relationship between the expectation of X_k^2 and $\sigma_x^2(k)$ can be described as

$$E(X_k^2) = 1/\left[1/\pm_x^2(k)\right] = \pm_x^2(k). \tag{7}$$

In a similar way

$$E\left(N_k^2\right) = 1/\left[1/\pm_n^2(k)\right] = \pm_n^2(k). \tag{8}$$

According to the Bayes' rule, the posteriori PDF of X_k^2 can be calculated as follows

$$f_{X_k^2}(X_k^2 \mid Y_k^2) = f_{Y_k^2}(Y_k^2 \mid X_k^2)\cdot f_{X_k^2}(X_k^2)/\, f_{Y_k^2}(Y_k^2), \tag{9}$$

$$f_{Y_k^2}\left(Y_k^2 \mid X_k^2\right) = f_{N_k^2}\left(Y_k^2 - X_k^2\right) =$$
$$= \left[1/\pm_n^2(k)\right]\cdot\exp\left[-\left(Y_k^2 - X_k^2\right)/\pm_n^2(k)\right], \tag{10}$$

$$f_{Y_k^2}\left(Y_k^2\right) = 1/\left[\pm_x^2(k) - \pm_n^2(k)\right],$$
$$\left\{\exp\left[-Y_k^2/\pm_x^2(k)\right] - \exp\left[-Y_k^2/\pm_n^2(k)\right]\right\},$$
$$if\ \pm_x^2(k) \neq \pm_n^2(k). \tag{11}$$

Before inserting (5), (10), (11) into (9), we define two intermediate variables:

$$1/\sim(k) = 1/\pm_x^2(k) - 1/\pm_n^2(k),\ if\ \pm_x^2(k) \neq \pm_n^2(k), \tag{12}$$

$$_k = \left[1/\sim(k)\right]/\left\{1 - \exp\left[-Y_k^2/\sim(k)\right]\right\}. \tag{13}$$

Then we get

$$f_{X_k^2}(X_k^2 \mid Y_k^2) = \ _k \exp[-X_k^2/\sim(k)],\ if\ \ \pm_x^2(k) \neq \pm_n^2(k). \tag{14}$$

When $\sigma_x^2(k)$ = $\sigma_n^2(k)$, we assume

$$f_{X_k^2}\left(X_k^2 \mid Y_k^2\right) = 1/Y_k^2,\ if\ \ \pm_x^2(k) = \pm_n^2(k). \tag{15}$$

Noticed that

$$\int_0^{Y_k^2}\left(1/Y_k^2\right)\cdot dX_k^2 = 1. \tag{16}$$

The assumption (15) is reasonable.

III. PROPOSED SPEECH ENHANCEMENT METHOD

As mentioned above, usage of IdBM can obtain high intelligibility, but the usage of this method needs accurate local SNR. In the meantime, the true local SNR is difficult to estimate in practice as we don't have enough information about clean sentences. The local SNR is defined as follows

$$\P_{L,k} = X_k^2 / N_k^2. \tag{17}$$

Following the approach in [16], the IdBM can be formulated using the following binary hypothesis model:

$$\begin{cases} H_1 : \P_{L,k} < thr, \\ H_2 : \P_{L,k} \geq thr. \end{cases} \tag{18}$$

Here *thr* represent the threshold of local SNR

$$\overline{X_k^2} = \begin{cases} Y_k^2, & when\ H_2\ is\ true, \\ 0, & when\ H_1\ is\ true. \end{cases} \tag{19}$$

According to (18), if we try to estimate $\overline{X_k^2}$ we need accurate local SNR which can be hardly archived without clean speech. There is another approach to get rid of this problem. Taking the expectation on $\overline{X_k^2}$, we get

$$E\left(\overline{X_k^2}\right) = E\left(G_k{}^2\right)\times Y_k^2 = Y_k^2 \times$$
$$\times\left[E\left(G_k{}^2 \mid H_1\right)\times P\left(H_1\right) + E\left(G_k{}^2 \mid H_2\right)P\left(H_2\right)\right], \tag{20}$$

where $P(H_x)$ represents the probability of hypothesis H_x is true, $E[G_k|H_x]$ denotes the gain function when hypothesis H_x is true. $P(H_2)$ is the key to this equation, its definition is the posteriori probability of $\xi_{L,k} > \theta$ as follows

$$P\left(H_2\right) = P\left(\P_{L,k} > thr \mid Y_k^2\right) = P\left(X_k^2 / N_k^2 > thr\right). \tag{21}$$

Insert (4), (17) into (21)

$$\begin{aligned} P(H_2)&=P\left[X_k^2>Y_k^2\cdot thr/(thr+1)\,|\,Y_k^2\right]= \\ &=\int_{Y_k^2\cdot thr/(thr+1)}^{Y_k^2} f_{X_k^2}\left(2\,|\,Y_k^2\right)d2. \end{aligned} \tag{22}$$

Insert (14) into (22)

$$\begin{aligned} P(H_2)&=\left\{\exp\left[\text{ª}_k/(thr+1)\right]-1\right\}/\left[\exp(\text{ª}_k)-1\right], \\ &\quad if\ \ \pm_x^2(k)\neq\pm_n^2(k). \end{aligned} \tag{23}$$

λ_k is an intermediate variable, defined as

$$\begin{aligned} \text{ª}_k&=\left[Y_k^2/\pm_n^2(k)\right]\times \\ &\times\left\{\left[1-\pm_x^2(k)/\pm_n^2(k)\right]/\pm_x^2(k)\times\pm_n^2(k)\right\}. \end{aligned} \tag{24}$$

Insert (15) into (21)

$$P(H_2)=1/(thr+1),\ if\ \pm_x^2(k)=\pm_n^2(k). \tag{25}$$

Notice that σ_x^2/σ_n^2 is exactly the definition of the *a priori* SNR ξ_k, we could use the "decision-directed" [2] approach to calculate it.

Then, we focus on the $E(G_k\,|H_1)$ and $E(G_k\,|H_2)$. For the former, $E(G_k\,|H_1) = 0$ would be reasonable as it is consistent with the IdBM method. As for $E(G_k\,|H_2)$, to compress residual noise, it should be less than one and consistent with the ξ_k. According to this principle, we choose two typical and classical forms, Wiener and Minimum Mean Square Error (MMSE) based on magnitude-squared spectrum [4].

$$E(G_k\,|\,H_2)=G_{Wiener}^2=\text{¶}_k/\left(\text{¶}_k+1\right), \tag{26}$$

$$\begin{aligned} E[G_k\,|\,H_2]&=G_{MMSE}^2= \\ &=\begin{cases} 1/2, & if\ \pm_x^2(k)=\pm_n^2(k), \\ 1/\text{ª}_k-1/\left[\exp(\text{ª}_k)-1\right], & else. \end{cases} \end{aligned} \tag{27}$$

Inserting (26), (27) into (20) respectively, we get:

$$\begin{aligned} G1_k&=\sqrt{X_k^2/Y_k^2}= \\ &=\begin{cases} \sqrt{\text{¶}_k/\left(\text{¶}_k+1\right)/(thr+1)}, & if\ \ \pm_x^2(k)=\pm_n^2(k), \\ \sqrt{\text{¶}_k/\left(\text{¶}_k+1\right)\cdot\left\{\exp\left[\text{ª}_k/(thr+1)\right]-1\right\}/\left[\exp(\text{ª}_k)-1\right]}, & else, \end{cases} \end{aligned} \tag{28}$$

$$\begin{aligned} G2_k&=\sqrt{X_k^2/Y_k^2}= \\ &=\begin{cases} \sqrt{\dfrac{1}{2}\cdot\dfrac{1}{thr+1}}, & if\ \ \pm_x^2(k)=\pm_n^2(k), \\ \sqrt{\dfrac{\left\{\exp\left[\text{ª}_k/(thr+1)\right]-1\right\}\cdot\left\{1/\text{ª}_k-1/\left[\exp(\text{ª}_k)-1\right]\right\}}{\exp(\text{ª}_k)-1}}, & else. \end{cases} \end{aligned} \tag{29}$$

We denote (28) as SMPO-Wiener and denote (29) as SMPO-MMSE. To compare these methods intuitively, we fix the *thr* at 0 dB and let the *a priori* SNR ξ range from -10 dB to 20 dB, then the corresponding gain functions are plotted in Fig. 1.

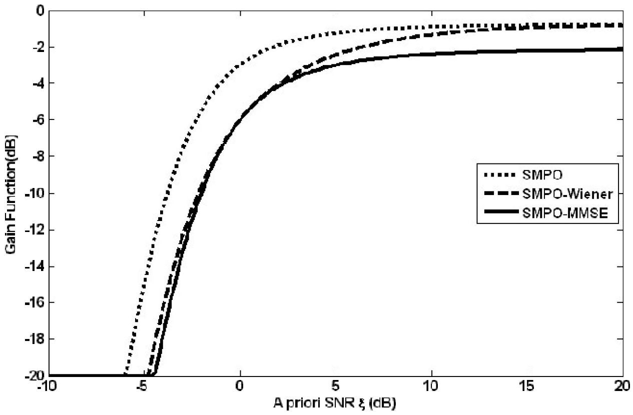

Fig. 1. Gain functions of the SMPO, SMPO-Wiener and SMPO-MMSE, respectively, as a function of the *a priori* SNR ¶. The threshold was fixed to *thr*=0 dB.

As can be seen from the figure, the gain functions of the SMPO-Wiener and the SMPO-MMSE are more aggressive than that of the SMPO when the *a priori* SNR ξ is low, which indicates better performance in noise reduction. When the *a priori* SNR ξ is high, the SMPO-Wiener and the the SMPO are almost the same, while the SMPO-MMSE is obviously low.

IV. EXPERIMENTS AND RESULTS

We choose the NOIZEUS [17] database to be our experimental corpus. This database has 30 clean English sentences, each of them has eight kinds of noisy sentences and four levels of SNR, which means 960 noisy sentences would be processed. The noise types include car, street, babble, exhibition, restaurant, station, train and airport. The four levels of SNR are 0 dB, 5 dB, 10 dB, 15 dB. After speech processing, we use two objective measures, i.e., the segmental SNR and the Perceptual Evaluation of Speech Quality (PESQ) [18], to assess effects of mentioned methods. Both measures are widely used in the speech enhancement area, besides, the PESQ has been proved to highly correlate with speech intelligibility. As a supplement, higher segmental SNR and PESQ values indicate better performance.

A. *Best SNR Threshold for Methods*

The threshold is important for the methods mentioned above; here we find the best thresholds for SMPO-Wiener and SMPO-MMSE firstly. As for the SMPO, the best threshold has been proved to be 0 dB by [4].

The research in [5] shown that the proper range of SNR threshold value should be [-12, 5] dB. In our experiments, the thresholds range from -10 dB to 5 dB, and the step is 5 dB. Here, we choose four kinds of noises, including babble noise, car noise, street noise, and airport noise. Each kind of noisy speech has four SNR levels, i.e., 0 dB, 5 dB, 10 dB, and 15 dB. We use SMPO-Wiener and SMPO-MMSE methods to process these degraded speech sentences respectively. After that, we compute the PESQ and segmental SNR values for each processed sentence, and then, we calculate statistical average values. The results are given in Table I and Table II, where Table I shows the segmental SNR results and the PESQ results are shown in Table II.

From Tables I and II, we find that both SMPO-Wiener and SMPO-MMSE methods perform best, in terms of PESQ values, when the threshold is -5 dB. After enough examining experiments, we found it consistent for all types of noise. As PESQ has a higher correlation with speech intelligibility than segmental SNR, we assumed -5 dB to be the best threshold for SMPO-Wiener and SMPO-MMSE methods and used it in further experiments.

TABLE I. EXPERIMENTAL RESULTS OF SMPO-WIENER AND SMPO-MMSE AS A FUNCTION OF THRESHOLD, *THR*, IN TERMS OF PESQ.

Noise	Method	*thr*	15 dB	10 dB	5 dB	0 dB
Car	SMPO -Wiener	-5dB	**3.039**	**2.690**	**2.304**	**1.985**
		0dB	2.967	2.622	2.238	1.902
		5dB	2.893	2.549	2.174	1.840
	SMPO -MMSE	-5dB	**3.005**	**2.662**	**2.279**	**1.959**
		0dB	2.937	2.592	2.206	1.878
		5dB	2.874	2.528	2.153	1.824
Babble	SMPO -Wiener	-5dB	**2.942**	**2.547**	**2.161**	1.796
		0dB	2.940	2.535	2.155	1.786
		5dB	2.926	2.516	2.138	**1.797**
	SMPO -MMSE	-5dB	**2.933**	**2.542**	**2.161**	1.793
		0dB	2.921	2.518	2.135	1.775
		5dB	2.907	2.498	2.117	**1.808**
Street	SMPO -Wiener	-5dB	**2.899**	**2.566**	**2.215**	**1.842**
		0dB	2.868	2.538	2.184	1.803
		5dB	2.820	2.511	2.154	1.785
	SMPO -MMSE	-5dB	**2.873**	**2.548**	**2.197**	**1.826**
		0dB	2.837	2.517	2.162	1.786
		5dB	2.796	2.493	2.134	1.767
Airport	SMPO -Wiener	-5dB	**2.937**	**2.564**	**2.205**	**1.805**
		0dB	2.922	2.542	2.181	1.774
		5dB	2.894	2.511	2.154	1.764
	SMPO -MMSE	-5dB	**2.928**	**2.552**	**2.194**	**1.791**
		0dB	2.904	2.523	2.161	1.759
		5dB	2.878	2.494	2.136	1.750

TABLE II. EXPERIMENTAL RESULTS OF SMPO-WIENER AND SMPO-MMSE AS A FUNCTION OF THRESHOLD, *THR*, IN TERMS OF SEGMENTAL SNR.

Noise	Method	*thr*	15 dB	10 dB	5 dB	0 dB
Car	SMPO -Wiener	-5dB	**7.587**	**4.610**	**1.812**	-2.786
		0dB	7.332	4.470	1.753	**-0.665**
		5dB	6.935	4.137	1.478	-0.871
	SMPO -MMSE	-5dB	**7.461**	**4.526**	**1.770**	**-0.656**
		0dB	7.222	4.370	1.657	-0.731
		5dB	6.876	4.083	1.429	-0.905
Babble	SMPO -Wiener	-5dB	**6.836**	3.663	0.510	-2.271
		0dB	6.831	**3.712**	0.660	-2.011
		5dB	6.679	3.616	**0.674**	**-1.863**
	SMPO -MMSE	-5dB	**6.779**	3.628	0.509	-2.239
		0dB	6.752	**3.636**	0.611	**-2.047**
		5dB	6.613	3.550	**0.621**	-2.415
Street	SMPO -Wiener	-5dB	**7.049**	**4.237**	1.135	-1.339
		0dB	6.884	4.219	**1.193**	-1.211
		5dB	6.593	4.082	1.127	**-1.195**
	SMPO -MMSE	-5dB	**6.948**	**4.169**	1.096	-1.336
		0dB	6.780	4.136	**1.128**	-1.263
		5dB	6.528	4.015	1.066	**-1.243**
Airport	SMPO -Wiener	-5dB	**7.013**	3.770	0.662	-1.913
		0dB	6.971	**3.819**	**0.758**	-1.681
		5dB	6.792	3.713	0.707	**-1.574**
	SMPO -MMSE	-5dB	**6.958**	3.729	0.650	-1.878
		0dB	6.887	**3.743**	**0.694**	-1.705
		5dB	6.731	3.657	0.655	**-1.616**

B. *Results and Comparison of Methods*

In the following experiments, the above-mentioned methods, i.e., SMPO, SMPO-Wiener, and SMPO-MMSE are applied to noisy sentence processing. After that, we calculate PESQ and segmental SNR values of processed sentences. Table III and Table IV show our statistical experiment results.

TABLE III. PERFORMANCE OF THE MENTIONED METHODS IN TERMS OF PESQ.

Noise	Method	15 dB	10 dB	5 dB	0 dB
Babble	Unprocessed	2.653	2.321	2.006	1.705
	SMPO	2.914	2.525	2.134	1.763
	SMPO-Wiener	**2.942**	**2.5467**	**2.161**	**1.796**
	SMPO-MMSE	2.933	2.542	2.161	1.793
Street	Unprocessed	2.541	2.247	1.904	1.563
	SMPO	2.835	2.517	2.163	1.785
	SMPO-Wiener	**2.899**	**2.566**	**2.215**	**1.842**
	SMPO-MMSE	2.873	2.548	2.197	1.826
Train	Unprocessed	2.492	2.160	1.859	1.605
	SMPO	2.905	2.496	2.137	1.771
	SMPO-Wiener	**2.936**	**2.522**	**2.164**	**1.806**
	SMPO-MMSE	2.924	2.518	2.155	1.792
Station	Unprocessed	2.578	2.249	1.959	1.665
	SMPO	2.948	2.606	2.283	1.851
	SMPO-Wiener	**2.983**	**2.627**	**2.300**	**1.865**
	SMPO-MMSE	2.950	2.610	2.284	1.851
Car	Unprocessed	2.532	2.201	1.891	1.634
	SMPO	3.010	2.668	2.283	1.951
	SMPO-Wiener	**3.039**	**2.690**	**2.304**	**1.985**
	SMPO-MMSE	3.005	2.662	2.279	1.959
Airport	Unprocessed	2.633	2.341	2.021	1.726
	SMPO	2.915	2.540	2.179	1.785
	SMPO-Wiener	**2.937**	**2.564**	**2.205**	**1.805**
	SMPO-MMSE	2.928	2.552	2.194	1.791
Exhibition	Unprocessed	2.514	2.184	1.882	1.585
	SMPO	2.879	2.526	2.133	1.678
	SMPO-Wiener	**2.908**	**2.550**	**2.154**	**1.701**
	SMPO-MMSE	2.888	2.535	2.143	1.686
Restaurant	Unprocessed	2.660	2.369	2.001	1.754
	SMPO	2.851	2.497	2.096	1.771
	SMPO-Wiener	**2.873**	**2.519**	**2.119**	**1.791**
	SMPO-MMSE	2.867	2.512	2.110	1.780

TABLE IV. PERFORMANCE OF THE MENTIONED METHODS IN TERMS OF SEGMENTAL SNR.

Noise	Method	15 dB	10 dB	5 dB	0 dB
Babble	Unprocessed	4.854	1.420	-1.783	-4.632
	SMPO	6.621	3.414	0.272	-2.515
	SMPO-Wiener	**6.836**	**3.663**	**0.510**	-2.271
	SMPO-MMSE	6.779	3.628	0.509	**-2.239**
Street	Unprocessed	5.196	1.843	-1.582	-4.258
	SMPO	6.747	4.114	1.116	**-1.263**
	SMPO-Wiener	**7.049**	**4.237**	**1.135**	-1.339
	SMPO-MMSE	6.948	4.169	1.096	-1.336
Train	Unprocessed	4.876	1.417	-1.691	-4.504
	SMPO	7.306	4.105	1.307	-1.442
	SMPO-Wiener	**7.510**	**4.354**	**1.535**	**-1.229**
	SMPO-MMSE	7.439	4.308	1.504	-1.231
Station	Unprocessed	4.715	1.235	-1.895	-4.712
	SMPO	7.038	3.917	1.140	-1.536
	SMPO-Wiener	**7.220**	**4.128**	**1.302**	**-1.365**
	SMPO-MMSE	7.086	4.052	1.247	-1.372
Car	Unprocessed	4.324	0.991	-2.173	-4.960
	SMPO	7.408	4.431	1.643	-0.748
	SMPO-Wiener	**7.587**	**4.610**	**1.812**	-0.694
	SMPO-MMSE	7.461	4.526	1.770	**-0.656**
Airport	Unprocessed	4.841	1.581	-1.672	-4.414
	SMPO	6.787	3.528	0.420	-2.138
	SMPO-Wiener	**7.013**	**3.770**	**0.662**	-1.913
	SMPO-MMSE	6.958	3.729	0.650	**-1.878**
Exhibition	Unprocessed	4.835	1.399	-1.838	-4.671
	SMPO	7.053	4.035	1.035	-1.623
	SMPO-Wiener	**7.230**	**4.242**	1.245	-1.426
	SMPO-MMSE	7.168	4.235	**1.275**	**-1.358**
Restaurant	Unprocessed	5.197	1.891	-1.390	-4.193
	SMPO	6.602	3.431	0.152	-2.511
	SMPO-Wiener	**6.783**	**3.676**	**0.413**	-2.271
	SMPO-MMSE	6.732	3.623	0.409	**-2.271**

From Table III and Table IV we notice that the SMPO-Wiener performs best, its segmental SNR improve significantly. All in all, our methods, both the SMPO-Wiener and the SMPO-MMSE, perform better than the original method.

Figure 2 shows the improvement in terms of PESQ when compared to unprocessed speeches. Figure 3 and Fig. 4 show timing waveforms and spectrograms of the speeches. Among the three methods, the SMPO-Wiener performed best. The reason why the SMPO-MMSE is not as good as the SMPO-Wiener can be seen from Fig. 1. When the *a priori* SNR ξ is high, the gain value of SMPO-MMSE is obviously lower than that of SMPO-Wiener, which means more speech distortion.

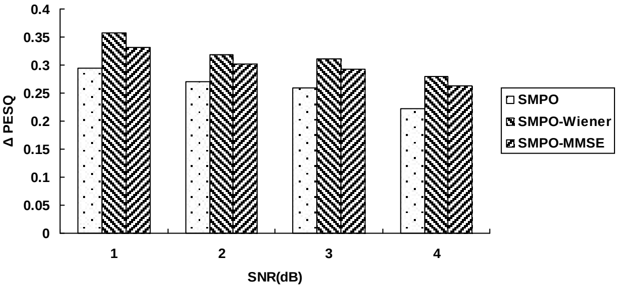

Fig. 2. Comparison of three methods in terms of ΔPESQ = PESQ_processed – PESQ_unprocessed, The speech was degraded by street noise. (1) SNR = 0 dB, (2) SNR = 5 dB, (3) SNR = 10 dB, (4) SNR = 15 dB.

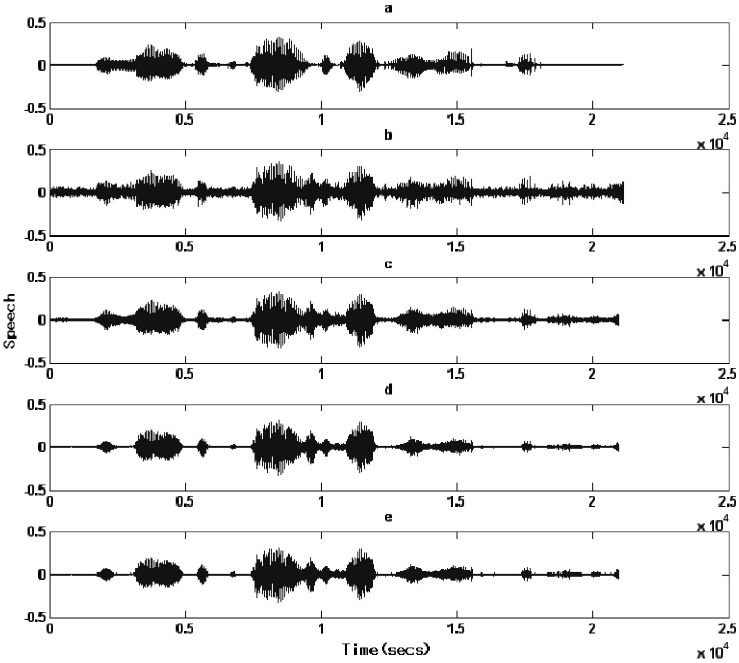

Fig. 3. Timing waveforms of: (a) clean speech signal, (b) noise corrupted speech signal with street noise 5 dB SNR and enhancement using, (c) SMPO, (d) SMPO-Wiener and, (e) SMPO-MMSE.

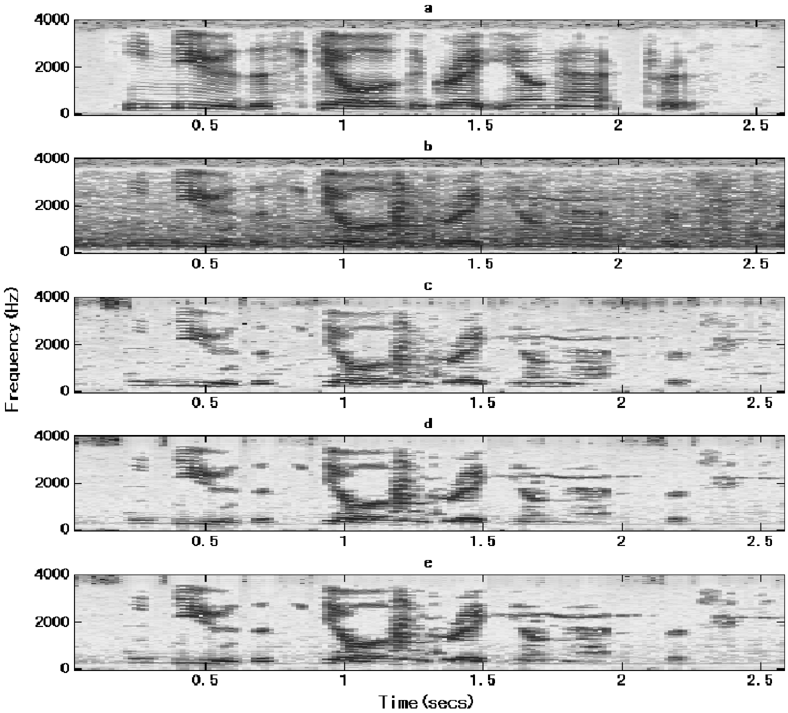

Fig. 4. Spectrograms of: (a) clean speech signal, (b) noise corrupted speech signal with street noise 5 dB SNR and enhancement using, (c) SMPO, (d) SMPO-Wiener and, (e) SMPO-MMSE.

V. Conclusions

In this paper, a new soft masking method was derived incorporating SNR uncertainty to enhance noisy speech. Compared to the conventional SMPO method, the proposed SMPO-Wiener and SMPO-MMSE methods yielded better performance owing to compressing residual noise. Comparing the SMPO-Wiener and the SMPO-MMSE, we analysed the reason why the SMPO-Wiener is more suitable than the SMPO-MMSE in this condition. Meanwhile, the difference between the SMPO-Wiener and the SMPO-MMSE means that there is still potential to improve performance by finding more suitable forms of $E(G_k|H_2)$. Besides, we realized that maybe we can change the binary masking model into other masking forms, because the noise would be totally masked by auditory masking effect when the local SNR value is high enough. In this condition, further compressing noise would be useless or even harmful for speech intelligibility and quality. In our future research, we would make efforts on these issues.

Appendix A

In this section, the PDF of Y_k^2 presented in (11) would be deduced. With the known condition $Y_k^2 = X_k^2 + N_k^2$, we can get the following equation

$$f_{Y_k^2}\left(Y_k^2\right)=\int_0^{Y_k^2} f_{X_k^2}\left({}^2\right) f_{N_k^2}\left(Y_k^2-{}^2\right)d^2, \tag{30}$$

where τ is an integral variable. Insert (5) and (6) into (30):

$$f_{Y_k^2}\left(Y_k^2\right)=1/\left[\pm_x^2(k)\cdot\pm_n^2(k)\right]\times$$
$$\times\int_0^{Y_k^2}\exp\left[-{}^2/\pm_x^2(k)\right]\exp\left[\left({}^2-Y_k^2\right)/\pm_n^2(k)\right]d^2= \tag{31}$$
$$=M\int_0^{Y_k^2}\exp\left\{{}^2\left[1/\pm_n^2(k)-1/\pm_x^2(k)\right]\right\}d^2,$$
$$M=\exp\left[-Y_k^2/\pm_n^2(k)\right]/\pm_x^2(k)/\pm_n^2(k). \tag{32}$$

After calculating the definite integral, we get

$$f_{Y_k^2}\left(Y_k^2\right)=N\times\exp\left\{{}^2\left[1/\pm_n^2(k)-1/\pm_x^2(k)\right]\right\}\Big|_0^{Y_k^2}=$$
$$=1/\left[\pm_x^2(k)-\pm_n^2(k)\right]\times \tag{33}$$
$$\times\left\{\exp\left[-Y_k^2/\pm_x^2(k)\right]-\exp\left[-Y_k^2/\pm_n^2(k)\right]\right\},$$
$$N=M/\left[1/\pm_n^2(k)-1/\pm_x^2(k)\right]. \tag{34}$$

Acknowledgment

I would like to express my gratitude to all those who helped me during the writing of this paper. A special acknowledgement should be shown to Gang-JinWang and Zhiqiang Lu for providing many useful suggestions about writing.

References

[1] D. Wang, G. Brown, *Computational auditory scene analysis: principles, algorithms, and applications.* Piscataway, NJ: Wiley/IEEE Press, 2006.

[2] Y. Lu, P. C. Loizou, "Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty", *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1123–1136, July 2011. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2082531

[3] Y. Hu, P. C. Loizou, "Evaluation of objective quality measures for speech enhancement", *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1109/TASL.2007.911054

[4] P. C. Loizou, G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions", *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 47–56, Jan., 2011. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2045180

[5] D. S. Brungart, P. S. Chang, B. D. Simpson, D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation", *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4007–4018, 2006. [Online]. Available: http://dx.doi.org/10.1121/1.2363929

[6] N. Li, P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction", *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673–1682, 2008. [Online]. Available: http://dx.doi.org/10.1121/1.2832617

[7] S. F. Boll, "Suppression of acoustic noise in speech using spectral sub-traction", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979. [Online]. Available: http://dx.doi.org/10.1109/TASSP.1979.1163209

[8] M. Berouti, M. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1979, pp. 208–211.

[9] W. Etter, G. S. Moschytz, "Noise reduction by noise-adaptive spec-tral magnitude expansion", *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.

[10] B. L. Sim, Y. C. Tong, J. S. Chang, C. T. Tan, "A parametric formulation of the generalized spectral subtraction method", *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, July 1998. [Online]. Available: http://dx.doi.org/10.1109/89.701361

[11] E. J. Diethorn, "Subband noise reduction methods for speech enhancement", in *Conf. on Acoustic Signal Processing for Telecommunication*, MA: Kluwer, 2000, pp. 155–178. [Online]. Available: http://dx.doi.org/10.1007/978-1-4419-8644-3_9

[12] C. Faller, J. Chen, "Suppressing acoustic echo in a spectral envelope space", *IEEE Trans. Speech Audio Process*, vol. 13, no. 5, pp. 1048–1062, Sept. 2005. [Online]. Available: http://dx.doi.org/10.1109/TSA.2005.852012

[13] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984. [Online]. Available: http://dx.doi.org/10.1109/TASSP.1984.1164453

[14] J. Jensen, I. Batina, R. C. Hendriks, R. Heusdens, "A study of the distribution of time-domain speech samples and discrete Fourier coefficients", in *Proc. SPS-DARTS*, 2005, vol. 1, pp. 155–158.

[15] A. Papoulis, S. U. Pillai, *Probability, Random Variables and Stochastic Processes.* 4th ed. New York: McGraw-Hill, 2002.

[16] R. McAulay, M. Malpass, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Trans. Acoust., Speech Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980. [Online]. Available: http://dx.doi.org/10.1109/TASSP.1980.1163394

[17] Y. Hu, P. C. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms", *Speech Commun.*, vol. 49, pp. 588–601, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2006.12.006

[18] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", 2000.