# Speech Enhancement Algorithm of Binary Mask Estimation Based on *a Priori* SNR Constraints

Jie Wang*, Chengcheng Yang*, Linhuang Yan*, Manlu Huang* and Jinqiu Sang†
*Guangzhou University, Guangzhou, China
E-mail: 2484888182@qq.com Tel: +86-20-39366923
† Communication Acoustics Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
E-mail: 464722621@qq.com Tel: +86-10-82547851

***Abstract*— A speech enhancement algorithm using binary mask estimation and *a priori* SNR constraints is proposed. The *a priori* SNR estimation has a major impact on noise spectrum estimation function, so the MMSE rule is used to modify the *a priori* SNR through secondary processing to get more accurate estimated noise power spectrum and gain function, which are used to retain noise over-estimated Time-Frequency units and discard noise under-estimated Time-Frequency units. Experiments results show that the proposed algorithm can improve the intelligibility of speech signals for low SNR.**

## I. INTRODUCTION

As an interdisciplinary of Linguistics and Signal processing, speech enhancement technology has made great progress with the development of computer and digital signal processing technology in recent years. Among numerous speech enhancement algorithms, for instance, the algorithms based on statistical model and adaptive filtering can suppress a large amount of background noise and then improve the quality of speech. However, it is less than desirable when these approaches improve the intelligibility of speech [1]. For this purpose, an ideal binary mask technique was designed to promote voice intelligibility by limiting speech distortion and imposing constraints on noise [2]. This method retains the time-frequency (T-F) regions where the target speech dominates the masker (noise) (e.g., local signal to noise ratio (SNR)>0 dB) and removes T-F units where the masker dominates (e.g., local SNR<0 dB). Therefore, combining the binary mask technique with a supervised learning approach, enhanced speech-noise energy ratio can be adaptively controlled. Furthermore, experimental findings by investigative analysis verified that the existing speech enhancement algorithms will bring about speech amplification distortion and attenuation distortion in the amendatory process, especially the amplification distortion (in excess of 6.02 dB) was most detrimental to speech intelligibility [5], which just matches certain academic theory about binary mask technology based on speech distortion control. Besides, Some scholars also put forward binary mask based on noise spectrum constraints through examining the effects of noise power spectrum density (NPSD) overestimation or underestimation to speech intelligibility [6][7]. And then the appropriate noise spectrum constraint is adopted to make the target signal more intelligible.

In this paper, the minimum mean square error (MMSE) method refined by Hendriks [8] are chose for speech enhancement because it have high accuracy but low complexity in the estimate of NPSD. According to the analysis and research of the proposed approach, MMSE of the discrete Fourier coefficient [9] and the Decision-Directed (DD) algorithm have been utilized in the course of NPSD estimate. It is well-known that the estimate with DD algorithm is biased. Hence, combining MMSE criterion, we make a secondary processing about *a priori* SNR to revise NPSD and gain function. Finally, the new binary mask is estimated precisely by the modified noise power spectrum and the speech power spectrum. Experiments results show that the proposed algorithm can improve the intelligibility of speech signals for low SNR.

## II. BINARY MASK ESTIMATION BASED ON *A PRIORI* SNR CONSTRAINTS

### *A. NPSD estimate and preliminary speech enhancement*

A Speech signal can be modeled as $y(n) = x(n) + d(n)$, where $y(n)$ is noisy speech signal and is obtained by adding clean speech signal $x(n)$ and additive noise signal $d(n)$. Transforming to frequency domain, we have:

$$Y(k,l) = X(k,l) + D(k,l), \quad (1)$$

where $X(k,l)$, $D(k,l)$ and $Y(k,l)$ correspond to the FFTs of $x(n)$, $d(n)$ and $y(n)$, respectively. $k$ and $l$ are, respectively frame number and frequency index.

To the best of our knowledge, as one of the key technologies of speech enhancement, the accuracy of noise estimation directly determines the performance of speech enhancement system. In order to meet the needs of realistic scene, performance of the enhanced algorithm should be excellent with good real-time and low complexity. For this reason, MMSE method was used for noise estimation that was refined by Hendriks [8]. The algorithm has better performance than the minimum-statistical algorithm based on tracking the minimum of short-term PSD estimate [10] and also better than the noise estimation algorithm based on minima controlled recursive averaging [11]. Implementation steps are expressed as follows:

1. Compute the minimum mean square error expectation of noise.

Due to the additive noise signal model, on the grounds of Bayes criterion and the given condition of noisy speech, the minimum mean square error expectation of noise can be written as:

$$\mathrm{E}\{D^2|Y\} = \frac{\int_0^{+\infty}\int_0^{2\pi} d^2 f_{Y|D,\Delta}(y|d,\delta)d\delta dd}{\int_0^{+\infty}\int_0^{2\pi} f_{Y|D,\Delta}(y|d,\delta)d\delta dd}. \quad (2)$$

It is assumed that the DFT of the speech and noise signal is subject to the complex Gaussian distribution. So we have:

$$f_{Y|D,\Delta}(y|d,\delta) = \frac{1}{\pi\sigma_x{}^2}\exp\left(\frac{2dr\cos(\delta-\theta)-r^2-d^2}{\sigma_x{}^2}\right) \quad (3)$$

and

$$f_{D,\Delta}(d,\delta) = \frac{d}{\pi\sigma_d{}^2}\exp\left(-\frac{d^2}{\sigma_d{}^2}\right). \quad (4)$$

Using (3) and (4) to compute (1), we get:

$$\mathrm{E}\{D^2|Y\} = \left(\frac{1}{(1+\xi)^2} + \frac{\xi}{(1+\xi)\gamma}\right)|Y|^2, \quad (5)$$

where $\xi$ and $\gamma$ denotes *a priori* SNR and *a posterior* SNR respectively, and is given by:

$$\xi(k,l) = \frac{|X(k,l)|^2}{|D(k,l)|^2} \quad (6)$$

and

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{|D(k,l)|^2}. \quad (7)$$

2. Assess *a priori* SNR and deviation compensation.

The *a priori* SNR of Maximum likelihood estimation algorithm can be expressed as:

$$\hat{\xi}_{ML}(k,l) = \max\{\gamma(k,l) - 1, \xi_{\min}\}, \quad (8)$$

where $\xi_{min}$ is the minimum constraint of the *a priori* SNR and its typical value is $-25$ dB. Therefore, $\mathrm{E}\{D^2|Y;\hat{\xi}(k,l)\}$ can be obtained by computing (5) with (8).

ML algorithm is a biased estimate. For the sake of exploring accurate estimation of NPSD, the deviation factor *B* is modeled as follows:

$$B = \frac{\sigma_d{}^2}{\mathrm{E}_Y\{\mathrm{E}\{D^2|y;\hat{\xi}\}\}} = \frac{\sigma_d{}^2}{\int_R\int_\Theta \mathrm{E}\{D^2|y;\hat{\xi}\}f_Y(y)rd\theta dr}, \quad (9)$$

where $\hat{\xi}$ can be estimated from DD algorithm [12]:

$$\hat{\xi}_{DD}(k,l) = \max\{\alpha\frac{|Y^2(k,l-1)|}{\sigma_d(k,l-1)} + (1-\alpha)\gamma(k,l) - 1, \xi_{\min}\}. \quad (10)$$

3. Compute noise power spectrum.

Using (5) to (8), noise power spectrum can be got by:

$$\hat{\sigma}_d^2(k,l) = \mathrm{E}\{D^2|y;\hat{\xi}(k,l)\}B\left(\hat{\xi}_{DD}(k,l)\right). \quad (11)$$

After smoothing, we have:

$$\hat{\sigma}_d^2(k,l) = \beta\hat{\sigma}_d^2(k,l-1) + (1-\beta)\hat{\sigma}_d^2(k,l), \quad (12)$$

where $\beta$ equals to 0.8.

In order to overcome the sudden change of environmental noise, the first five frames of noise power spectrum are computed as minimum. So we finally get:

$$\hat{\sigma}_d(k,l) = \max\{\hat{\sigma}_d(k,l), P_{\min}(k,l)\}. \quad (13)$$

It is worth noting that the construct of deviation factor adopts DD algorithm to amend *a priori* SNR estimation. While DD algorithm has a frame delay, which causes that the initial enhancement speech contains incompletely suppressed background noise, on the other hand, some "music noise" will be introduced. Consequently, we will analyze the relation between *a priori* SNR estimation and noise power spectrum

estimation to address the problem in next section.

*B. Influence of a priori SNR estimation on noise function*

In view of the assumed model in section 2.1, estimate of speech and noise magnitude spectrum can be formulated by:

$$\hat{X}(k,l) = G_x(k,l) \times Y(k,l) \tag{14}$$

and

$$\hat{D}(k,l) = G_d(k,l) \times Y(k,l). \tag{15}$$

Benefit from a simple structure and easy to implement, DD algorithm are selected to derive noise spectrum gain function $G_d(k,l)$:

$$G_d(k,l) = \frac{1}{1+\xi(k,l)}\ , \tag{16}$$

where $\xi(k,l)$ denotes *a priori* SNR of speech signal at each frequency bin *k* and *l*.

It is assumed that $\xi' = \xi + \Delta\xi$ means *a priori* SNR estimate value with a estimate error $\Delta\xi$, and $\xi$ is real *a priori* SNR estimator. Then the difference of noise spectrum gain function $\Delta G_d(k,l)$ can be calculated from (16):

$$\Delta G_d(k,l) = \frac{1}{1+\xi'(k,l)} - \frac{1}{1+\xi(k,l)}$$

$$= \frac{\xi(k,l)-\xi'(k,l)}{(1+\xi'(k,l))(1+\xi(k,l))} = -\frac{\Delta\xi}{\left((1+\xi'(k,l)\right)(1+\xi(k,l))}. \tag{17}$$

So as to further reveal the correlation between *a priori* SNR and noise spectrum estimation, the error $\Delta\xi$ ranges from −30 to 30 dB. And then we obtain the relation between SNR estimate error and the difference of noise spectrum gain function in Fig. 1.

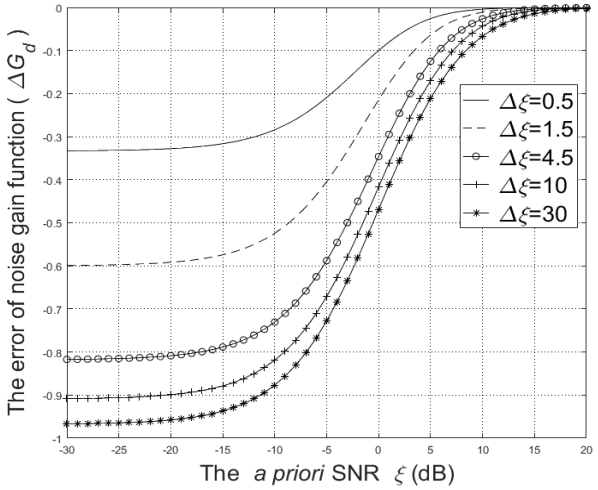


Figure.1 Influence of prior SNR estimation on noise spectrum gain function.

As can be seen from Fig. 1, the overestimation of *a priori* SNR will lead to the underestimation of NPSD. Considering that the minimum mean square error of DFT coefficient has been utilized in the course of NPSD estimate, hence, combing MMSE criterion, we make a correction again about *a priori* SNR estimate to achieve more accurate initial speech signal and NPSD. By utilizing human-being auditory masking effect and cochlear working mechanism, we make a binary masking about initial enhanced speech and noise signal in time-frequency domain to achieve a further improvement of speech quality.

The secondary processing focuses on *a priori* SNR estimate of DD algorithm in (10), given by [13]:

$$\xi_1(k,l) = \sqrt{\frac{\xi_{DD}(k,l)}{1+\xi_{DD}(k,l)}} \cdot \left(1 + \sqrt{\frac{\xi_{DD}(k,l)}{1+\xi_{DD}(k,l)}} \cdot \gamma(k,l)\right). \tag{18}$$

Applying the new *a priori* SNR to the proposed course of noise evaluation, the revised NPSD is retrieved. So the gain function of the proposed algorithm is expressed as:

$$G_1(k,l) = \frac{\xi_1(k,l)}{1+\xi_1(k,l)} \tag{19}$$

and

$$G_2(k,l) = \max(G_1, G_{\min}), \tag{20}$$

where $G_{min} = \text{eps}$ is a constraint value of gain function.

Therefore, we get the initial enhanced speech magnitude spectrum as:

$$\hat{X}_1(k,l) = G_2(k,l) \times Y(k,l). \tag{21}$$

*C. Construction of binary mask*

The proposed algorithm applies the ideal binary mask technique into speech enhancement by retaining the time-frequency (T-F) units of the mixture signal that are stronger than the interfering noise (masker), and removing the T-F units where the interfering noise dominates. Specifically, we simulate the characteristics of the basilar membrane of the cochlea, and design a gammatone filter bank whose center frequency is distributed with quasi logarithmic form, and whose frequency band ranges from 80 Hz to 5000 Hz. With

the gammatone filter bank, the initial speech signal and noise signal are transformed into two-dimensional representation in frequency-domain to compute frequency power at each time unit. Finally a new binary mask is obtained. The impulse response of gammatone filter can be expressed as [14]:

$$g(f,t)=\begin{cases} b^a t^{a-1} e^{-2\pi b(f)t}\cos(2\pi f t), \text{if } t\geq 0 \\ 0, \text{else} \end{cases}, \qquad (22)$$

where $a$ denotes filter order, and equals to 4. And $b(f)$ denotes equal rectangular bandwidth when center frequency is $f$.

So we get a further processing with initial speech signal through gammatone filter bank as [15]:

$$\hat{X}_2(k,l)=\left(\frac{1}{C}\right)\sum_{c=0}^{C-1}\left(\hat{X}_1(k,l)\times|G_k(c)|^2\right), \qquad (23)$$

where $C$ means DFT coefficient of speech signal. $G_k$ means filter frequency-response of group $k$ and is derived from (22).

Using (22) for estimating noise signal, so the spectrum estimation after the binary mask transformation can be written as:

$$\hat{X}_M(k,l)=\begin{cases} \hat{X}_2(k,l), \text{if } \hat{E}_X(k,l)-\hat{E}_d(k,l)\geq LC \\ 0, \text{else} \end{cases}, \qquad (24)$$

where $\hat{E}_X(k,l)$ and $\hat{E}_d(k,l)$ are, respectively, short-term energy of initial speech signal and noise signal, which both are obtained from preproccess. $LC$ is a threshold of *a priori* SNR estimator.

Fig. 2 shows the block diagram of the proposed binary mask estimation based on the *a priori* SNR constraints.

## III. Experimental Simulation And Analysis

As a test sample, clean speech were stitched together by many speakers' recording. The speakers wore headphone in a closed laboratory, and the distance from microphone to his sound source is about 2 cm. Noise samples were composed of white noise from Noisex-92 noise library, babble noise, factory noise and F16 noise. All sampling rate is 16 kHz. The noisy speech whose SNR ranges from 0 to 10 dB was generated by adding four kinds of noise in varying proportions into clean speech. The other parameters the proposed algorithm concerned include $\alpha$=0.98, $\beta$=0.8. Here, we regard Speech Presence Probabilities algorithm (SPP) by

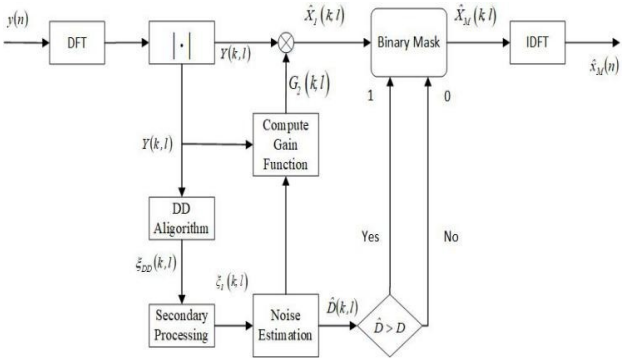


Fig. 2 Block diagram of the proposed binary mask estimation based on the *a priori* SNR constraints.

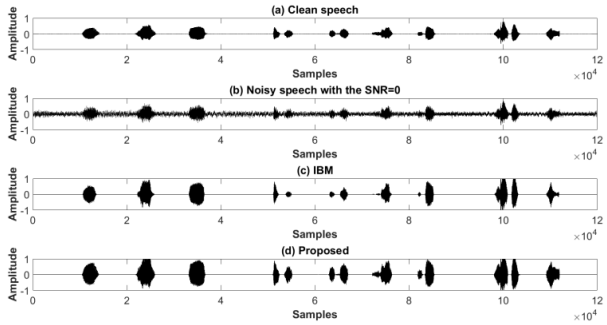


Fig. 3 Speech waveforms processed by IBM algorithm and the proposed method with 0 dB white noise.

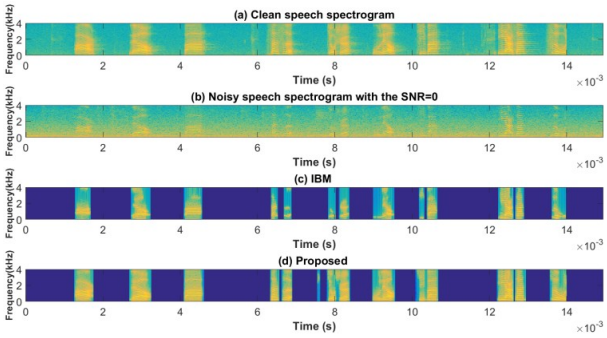


Fig. 4 Spectrograms processed by IBM algorithm and the proposed method with 0 dB white noise.

Zheng [15] as M1, regard IBM algorithm as M2 [16], regard our proposed algorithm as M3.

Fig. 3 and Fig. 4 illustrate, respectively, time-domain waveforms and spectrograms of enhanced speech processed by IBM algorithm and the proposed method with 0 dB white noise. Color intensity of speech in spectrograms represents how much speech there is, and the more concentrated the

color is, the more content the speech have, the more intelligible the speech are. Comparing processed spectrogram of IBM with the proposed method shows that though speech tone of IBM are more focus, it is somewhat ambiguous due to the amplification distortion. While the speech tone processed by our approach is more distinct then the result processed by IBM. This indicates the proposed algorithm can improve speech intelligibility obviously.

To further verify the effectiveness of the proposed algorithm, we made some subjective and objective test. Speech enhancement algorithm should improve the intelligibility of speech as much as possible when it can suppress noise. In auditory subjective test, assessment criteria is based on Mean Opinion Score (MOS) [17], while objective evaluation standard is depend on the average of segmental SNR improvement (SegSNRI) [18] and Average Logarithmic Spectrum Distance (LSD) [18].

In the test of MOS, 7 men and 3 women aged 22-25 with normal hearing were gathered to grade the speech data prepared before. The main basis is whether they can hear and understood. Priori to the sentence test, each subject listened to a set of noise-corrupted sentences to get familiar with the testing procedure. In the formal experiment, the subjects were asked to listen at random to the 30 groups of speech that would be unprocessed or processed by the proposed algorithm, and recorded the scores of each condition. Then the MOS can be calculated.

Fig. 5 shows evidently that under different noise enhancement algorithm remain some "music noise" and background noise more or less. While our proposed algorithm solves this problem well, and improves the articulation and intelligibility of speech.

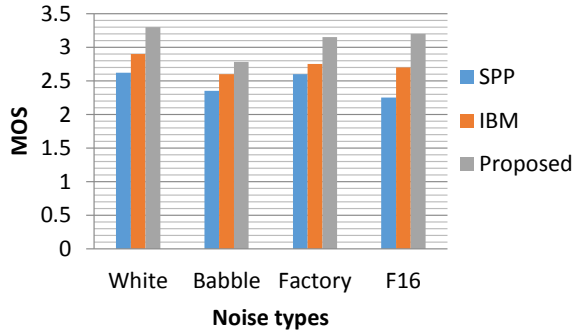


Fig. 5 MOS scores of three speech enhancement algorithms

Table 1. Comparisons of segmental SNR improvement for enhanced speech in various noise corruptions.

| Noise types | SNR(dB) | Method | | |
|---|---|---|---|---|
| | | M1 | M2 | M3 |
| White | 0 | 7.57 | 7.39 | **7.95** |
| | 5 | 5.76 | 6.05 | **6.40** |
| | 10 | 4.25 | 3.96 | **4.41** |
| Babble | 0 | 5.27 | 5.06 | **5.35** |
| | 5 | 3.14 | 3.67 | **3.99** |
| | 10 | 2.02 | 2.75 | **3.20** |
| Factory | 0 | **8.09** | 7.10 | 7.66 |
| | 5 | 7.25 | 6.71 | **7.33** |
| | 10 | 6.67 | 6.32 | **6.83** |
| F16 | 0 | 6.07 | 5.74 | **6.23** |
| | 5 | 4.41 | 4.42 | **5.03** |
| | 10 | 3.31 | 3.33 | **3.88** |

Table 2. Comparisons of LSD for the enhanced speech in various noise corruptions.

| Noise types | SNR(dB) | LSD | | |
|---|---|---|---|---|
| | | M1 | M2 | M3 |
| White | 0 | 5.74 | 5.88 | **5.63** |
| | 5 | 5.15 | 5.13 | **5.00** |
| | 10 | 4.48 | 4.54 | **4.31** |
| Babble | 0 | 5.02 | 4.51 | **3.75** |
| | 5 | 4.33 | 4.32 | **4.10** |
| | 10 | 3.67 | 3.23 | **3.02** |
| Factory | 0 | 4.35 | 4.57 | **4.22** |
| | 5 | 3.19 | 3.29 | **3.04** |
| | 10 | 2.30 | 2.35 | **2.18** |
| F16 | 0 | 5.58 | 5.72 | **5.43** |
| | 5 | 4.91 | 4.83 | **4.46** |
| | 10 | 3.94 | 3.69 | **3.42** |

Table 1 and Table 2 shows the SegSNRI and LSD comparison results of three speech enhancement algorithms. It is obvious that under any condition (white noise, babble noise, factory noise or F16 noise), the performance of new binary mask estimation based on the *a priori* SNR is better than the other two speech enhancement algorithms. Therefore,

according to the subjective and objective experiments, it is clear that the algorithm proposed in this paper can eliminate the background noise sufficiently and improve the speech quality.

## IV. CONCLUSIONS

This paper mainly researched the speech enhancement algorithm based on ideal binary mask. Firstly, the influence of *a priori* SNR overestimation or underestimation on NPSD was analyzed and studied. And it was discovered that the noise estimate method based MMSE have high accuracy but low calculation complexity. So MMSE was applied into ideal binary masking speech enhancement algorithm. Moreover, concretely analyzing the calculation process of noise estimation based on MMSE, it could be found that noise estimate was obtained by minimum mean square error estimator in the assumption that speech and noise are independent each other and both obeyed the Gaussian distribution. Hence, given that the estimate with DD algorithm is biased, we made a secondary processing based on MMSE about *a priori* SNR estimated in DD algorithm. And combining human-being auditory masking effect and cochlear working mechanism into ideal binary mask technique, the proposed algorithm has achieved pretty good results.

## ACKNOWLEDGMENT

This work was supported by National Science Fund of China (No.61302126), Special Innovation Project of Department of Education of Guangdong Province (No. 2017KTSCX141), Key Lab of Information Processing & Transmission of Guangzhou (No.201605030014）and Modern Video & Audio Information Engineering Center of Guangdong Province.

## REFERENCES

[1] J. Benesty, S. Makino, J. Chen, *Speech enhancement*, Springer, 2005.

[2] Li. N，Bao. C. C，Xia. B. Y and Bao. F, "Speech intelligibility improvement using the constraints on speech distortion and noise over-estimation," *IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, *IEEE Computer Society*, Beijing. pp. 602-606, 2013.

[3] Li. N and Loizou. P. C, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The Journal of the Acoustical Society of America,* vol. 123, no. 3, pp. 1673-1682, 2008.

[4] Djendi. M and Scalart. P, "Reducing over-and under-estimation of the *a priori* SNR in speech enhancement techniques," *Digital Signal Processing,* vol. 32, no. 2, pp. 124-136, 2014.

[5] Loizou. P. C and Kim. G, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing,* vol.19, no.1, pp. 47-56, 2011.

[6] Kim. G, "Binary Mask Criteria Based on Distortion Constraints Induced by a Gain Function for Speech Enhancement," *IEEE Transactions on Smart Processing and Computing*, vol.2, no. 4, pp. 197-202, 2013.

[7] Kim. G， Loizou. P. C, "A new binary mask based on noise constraints for improved speech intelligibility," *Interspeech,* Chiba, Japan. pp. 1632-1635, 2010.

[8] Hendriks. R.C, Heusdens. R and Jensen. J, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE ICASSP,* pp. 4266–4269, 2010.

[9] Erkelens. J. S, Hendriks. R. C, Heusdens. R, and Jensen. J, "Minimum mean square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process,* vol. 15, no. 6, pp. 1741-1752, 2007.

[10] Martin. R, " Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans Speech & Audio Processing*, vol. 9, pp. 504 – 512, Jul. 2001.

[11] Cohen. I and Berdugo. B, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters,* vol. 9, no. 1, pp. 12-15, 2002.

[12] Ephraim. Y and Malah. D, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics., Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, 1984.

[13] Alam. M. J, O'Shaughnessy. D. D and Selouani. S. A, "Speech enhancement based on novel two-step *a priori* SNR estimators," *Interspeech*, Brisbane, Australia. pp. 565-568, 2008.

[14] Patterson. R. D, Nimmo-Smith. I, Holdsworth. J and Rice. P,

"An efficient auditory filterbank based on the gammatone function," *A meeting of the IOC Speech Group on Auditory Modelling at RSRE*. 7 Feb. 1987.

[15] Zheng. C S, "A Modified *a Priori* SNR Estimator Based on the United Speech Presence Probabilities, " *Journal of Electronics & Information Technology*, vol. 30, no. 7, pp. 1680-1683, 2008.

[16] Narayanan. A and Wang. D. L, "A CASA-Based System for Long-Term SNR Estimation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 9, pp. 2518-2527, 2012.

[17] Hu. Y and Loizou. P. C, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229-238, 2008.

[18] Hu. Y and Loizou. P. C, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 57, pp. 214-231, 2008.