



An Optimal Speech Enhancement under Speech Uncertainty Probability and Masking Property of Auditory System

Xiaoshan Huang¹, Xiaoqun Zhao²

¹ Electronics and Information Institute of Tongji University, Shanghai, China

² Electronics and Information Institute of Tongji University, Shanghai, China

yellowsmallhill@hotmail.com, zhao_xiaoqun@mail.tongji.edu.cn

Abstract

Recently, I.Cohen has presented causal and noncausal algorithms to modify the classic decision-directed approach for prior SNR. It is well-known that prior SNR is critical to trade off the musical noise level and the audible clearness level in spectral subtraction speech enhancement. However, all these algorithms conflict with statistical signal model more or less. To adjust smoothing parameters which play an important role on the recursive procedure of prior SNR and noise spectrum estimate more reasonably, we present novel speech uncertainty state model which capitalizes on the masking property of auditory system, and propose a new modified approach which employs speech uncertainty probability to make automatic adaptation of smoothing parameters. Novel algorithm is capable of eliminating musical noise meanwhile lowering speech distortion by remaining original speech in the case of inaudible noise under masking threshold. Experiments confirm that novel algorithm is superior to classic methods, particularly at low SNR environment.

Index Terms: speech enhancement, speech uncertainty probability, auditory masking property

1. Introduction

Some isolated peaks are remained in spectral domain after the short-time spectral subtraction which known as the “musical noise” in perception. The annoying artifact in spectral subtraction is the universal phenomenon need to be solved. A lot of modifications of the basic suppression rules have been proposed in order to overcome the musical noise, but these techniques only reduce the musical noise partly at the sacrifice of the audible clearness. The approaches presented by Ephraim and Malah (which will be abbreviated as MMSE and MMLS in the following) counter the musical noise very well [1-2].Capper showed that the nonlinear smoothing procedure of the priori SNR named decision-directed approach in MMSE and MMLS makes them possible to obtain significant noise reduction while avoiding the musical noise described above. Furthermore, he analyzed the value of smoothing parameter in prior SNR qualitatively, which is the dominant factor to influence the degree of fluctuation of priori SNR in noisy area and the level of transient distortion brought to signal [3].However, the value of smoothing parameter is fixed empirically by Ephraim. In causal and noncausal algorithms, I.Cohen has set smoothing parameter time-varying by the statistical model. Unfortunately, the variable smoothing parameter is inconsistent with statistical model strictly. Consequently, these methods couldn’t achieve optimal trade-off between residual musical level and speech distortion level.

We focus here on MMLS speech enhancement system which minimizes mean-square error of log-spectral amplitude in its perception. It was verified that MMLS estimator is the optimal spectral estimator compared with another estimators, such as Weiner estimator, MMSE estimator, for the reason that MMLS estimator is more subjectively meaningful one for distortion measure [4-5]. Incorporating speech uncertainty probability initially proposed by Mcaclay [6], I.Cohen has presented optimally modified MMLS (for short OM-MMLS) estimator proved more efficient in reducing musical noise [5]. Taking the masking property of auditory system into account, we propose a new four-state speech uncertainty model rather than two-state. The original two-state hypotheses of speech absent H_0 and speech present H_1 are further divided into sub-two-state as noise masked and noise unmasked, associated with masking threshold. Then, gain and noise spectrum, prior SNR are derived under novel hypotheses.

This paper is organized as follows. In Section 2, we analyze estimator systems proposed by Ephraim and extended by I.Cohen to obtain major factors affecting the tradeoff between noise reduction and audible distortion. Furthermore, we formulate the problem in these estimator systems. In Section 3, we introduce new four-state speech uncertainty model in view of the masking property of auditory system and present new algorithm for gain, prior SNR, noise spectrum estimators under new assumption. In Section 4, we evaluate novel algorithm and discuss results of experiment, which validate its usefulness. In Section 5, we draw the conclusion for the whole estimator system.

2. Description of the problem

MMLS estimator system comprises three main parts, including gain function, noise spectrum estimate and prior SNR estimate. Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively, where n is a discrete-time index. Applying the short-time Fourier transform (STFT) to the observed signal $y(n)$, we have in the time-frequency domain

$$Y_{(l,k)} = X_{(l,k)} + D_{(l,k)} \tag{1}$$

where k is frequency-bin index and l is time frame index.

We adopt MMLS estimator’s gain G_{LS} as elementary gain for its superiority in perception related as above.

$$G_{LS(l,k)} = \frac{\xi_{(l,k)}}{1 + \xi_{(l,k)}} \exp \left\{ \frac{1}{2} \int_{v_{(l,k)}}^{\infty} \frac{e^{-t}}{t} dt \right\} \tag{2}$$

where $v_{(l,k)}$ could be expressed as below

$$v_{(l,k)} \triangleq \gamma_{(l,k)} \xi_{(l,k)} / (1 + \xi_{(l,k)}) \tag{3}$$

where $\xi_{(l,k)} \triangleq \lambda_{X(l,k)} / \lambda_{D(l,k)}$ and $\gamma_{(l,k)} \triangleq |Y_{(l,k)}|^2 / \lambda_{D(l,k)}$ represent the priori SNR and the posteriori SNR respectively, $\lambda_{X(l,k)} \triangleq E\{|X_{(l,k)}|^2\}$ and $\lambda_{D(l,k)} \triangleq E\{|D_{(l,k)}|^2\}$ denote respectively the short-term spectrum of speech and noise spectral signals.

We investigate here the IMCRA method proposed by I.Cohen to estimate $\lambda_{D(l,k)}$, because IMCRA carries out the smoothing of noisy power spectrum in both time and frequency given by averaging past spectral power values and takes into account two-state speech uncertainty (H_0 and H_1) rather than restricting the update of noise estimator to speech absence periods [7]. $\lambda_{D(l,k)}$ is written as

$$\begin{aligned} \lambda_{D(l,k)} &= \lambda_{D(l-1,k)} p_{1(l,k)} + \left[\alpha_D \lambda_{D(l-1,k)} + (1 - \alpha_D) |Y_{(l,k)}|^2 \right] p_{0(l,k)} \\ &= \hat{\alpha}_{D(l,k)} \lambda_{D(l-1,k)} + \left[1 - \hat{\alpha}_{D(l,k)} \right] |Y_{(l,k)}|^2 \end{aligned} \quad (4)$$

where $p_{1(l,k)} \triangleq P(H_1 | Y_{(l,k)})$, $p_{0(l,k)} \triangleq P(H_0 | Y_{(l,k)}) = 1 - p_{1(l,k)}$ represent speech present probability and speech absent probability respectively. $\hat{\alpha}_{D(l,k)} = \alpha_D + [1 - \alpha_D] p_{1(l,k)}$ is a time-varying smoothing parameter which is adjusted by the speech present probability. We here assume knowledge of speech uncertainty probability, which in practice can be estimated by method presented by I.Cohen [7].

Ephraim and Malah presented decision-directed approach (DD) to estimate ξ [1]. I.Cohen indicated that DD method doesn't agree with statistical signal model and introduced novel statistical model that takes into account the time-correlation between successive speech spectral components. Using information from neighboring frames and DD recursive procedure, he presented two new recursive estimators for priori SNR, including noncasual and casual estimators [8-10]. No matter how complicated those recursive procedures are, the fundamental mechanism of these three methods is similar. The basic formulation,

$$\xi_{(l,k)} = \tilde{\alpha}_{(l,k)} A_{(l,k)}^2 / \lambda_{D(l-1,k)} + (1 - \tilde{\alpha}_{(l,k)}) Q[\gamma_{(l,k)} - 1] \quad (5)$$

where $A_{(l,k)} \triangleq G_{LS} |Y_{(l,k)}|$ denotes the spectral speech amplitude, $Q[\bullet] = \max(\bullet, 0)$. The first term and the second term of (5) represent previous prior SNR and current frame's posteriori SNR respectively.

The rule of smoothing parameter $\tilde{\alpha}_{(l,k)}$ has been manifested: larger $\tilde{\alpha}_{(l,k)}$ is appropriately applied to speech state H_0 in which we mainly focus on musical noise reduction. In this case, previous prior SNR takes more important role to decide the value of current frame's prior SNR. Then, that results in less variance of prior SNR but slower feedback to the change of signal. Less variance of prior SNR provides a great reduction of musical noise, but more speech distortion. Reversely smaller value of $\tilde{\alpha}_{(l,k)}$ is fit for speech state H_1 in which we pay attention to speech distortion as well. Because the second term takes a more important role to decide current frame's prior SNR and musical noise is probably contributed to by the second term. In this case, ξ responding fast to the input signal reduces the distortions of the enhanced speech but introduces some residual "musical noise".

The theoretical investigation of the basic estimator is complicated due to its highly nonlinear nature. Therefore, in case of DD, $\tilde{\alpha}_{(l,k)}$ is fixed value by simulation only. Comparatively, we could simplify recursive procedure of prior SNR in casual and noncasual to basic formulation under special condition. $\tilde{\alpha}_{(l,k)}$ of these two estimators is monotonically decreasing as $\hat{A}_{(l-1,k)} / \lambda_{D(l-1,k)}$. Obviously, $\tilde{\alpha}_{(l,k)}$ in noncasual and casual recursive estimators conform to above desirable behavior qualitatively. However the adjusting of $\tilde{\alpha}_{(l,k)}$ is not quantitatively reasonable due to the conflict in the derivation of ξ . The update step of ξ in Casual and Noncasual estimators is derived according to the assumed statistical signal model. Unfortunately the derivation has conflict with the model. According to the derivation, wrote by I.Cohen [8]. The estimate for $\xi_{(l,k)}$ achieved given by

$$E\{X_{(l,k)} | \lambda_{X\text{pro}(l,k)}, Y_{(l,k)}\} = \frac{\lambda_{X\text{pro}(l,k)}}{\lambda_{X\text{pro}(l,k)} + \lambda_{D(l,k)}} Y_{(l,k)} \quad (6)$$

where $\lambda_{X\text{pro}(l,k)}$ denotes the variance of previous frames' speech spectrum. And we focus on the fourth characteristic of novel speech statistical model proposed by I.Cohen [8]. It is assumed that a spectral component $X_{(l,k)}$ is a zero-mean complex Gaussian random variable with independent and identical distribution (IID) real and imaginary parts. It is obvious that the assumption conflicts with expression (6).

Experimental results show that casual and noncasual are not better than OM-MMLS, which adopts the same noise spectral estimate, signal model, and the speech uncertainty probability as casual and noncasual, except adopting DD to estimate prior SNR. However these three estimators yield better results than the classic MMLS. It is proved that IMCRA method to estimate the noise spectrum continuously ensures them better than classic MMLS which uses the discontinuous noise spectrum estimate. As (4) implies, smoothing parameter $\hat{\alpha}_{D(l,k)}$ of noise spectrum in IMCRA is proportional to the speech presence probability, then noise spectrum could change fast to follow the input noise signal under H_0 and change slowly to reduce the variance of noise spectrum under H_1 . Additionally, Virag selected masking threshold as subtraction parameter in general spectral subtraction [11]. We introduce these two conceptions into prior SNR and noise spectrum estimate.

3. New method

In this paper, we integrate noise masked probability into speech uncertainty probability to adjust smoothing parameters. In view of masking property, the two-state speech uncertainty state model, including speech absent H_0 and speech present H_1 , will be evolved into four-state hypotheses model. The best solution to choose the enhancement system for each original state is based on the following consideration: if the masking threshold is high, residual noise will be naturally masked and inaudible. Hence, there is no need to reduce noise in order to keep distortion as low as possible. In this case, gain of estimator just is set as 1, meanwhile noise spectrum estimate and prior SNR estimate follow the input signal fast without any audible fluctuation. However, if masking threshold is too low to completely mask the residual noise, residual noise will be annoying to the listener and it is

necessary to reduce it. In this case, the gain of estimator need adopt MMLS estimator, similarly, noise spectrum and prior SNR estimate are deduced by DD approach to lower fluctuation of λ_D and ξ . The whole estimation system will be addressed under the four-state. Log spectrum gain, noise spectrum and prior SNR estimates, probability of each state are obtained as below.

$H_{0,0}$: Speech absent and noise unmasked state

$$\begin{aligned} G_{H_{0,0}} &= G_{\min}; \lambda_{D(l,k)} = \alpha_D \lambda_{D(l-1,k)} + (1-\alpha_D) |Y_{(l,k)}|^2; \\ \xi_{(l,k)} &= A_{(l-1,k)}^2 / \lambda_{D(l-1,k)}; p_{0,0(l,k)} = p_{0(l,k)} (1-p_{T(l,k)}) \end{aligned} \quad (7)$$

$H_{0,1}$: Speech absent and noise masked state

$$\begin{aligned} G_{H_{0,1}} &= 1; \lambda_{D(l,k)} = |Y_{(l,k)}|^2; \\ \xi_{(l,k)} &= A_{(l-1,k)}^2 / \lambda_{D(l-1,k)}; p_{0,1(l,k)} = p_{0(l,k)} p_{T(l,k)} \end{aligned} \quad (8)$$

$H_{1,0}$: Speech present and noise unmasked state

$$\begin{aligned} G_{H_{1,0}} &= G_{LS}; \xi_{(l,k)} = \alpha A_{(l-1,k)}^2 / \lambda_{D(l-1,k)} + (1-\alpha) Q[\gamma_{(l,k)} - 1]; \\ \lambda_{D(l,k)} &= \lambda_{D(l-1,k)}; p_{1,0(l,k)} = p_{1(l,k)} (1-p_{T(l,k)}) \end{aligned} \quad (9)$$

$H_{1,1}$: Speech present and noise masked state

$$\begin{aligned} G_{H_{1,1}} &= 1; \lambda_{D(l,k)} = \lambda_{D(l-1,k)}; \\ \xi_{(l,k)} &= Q[\gamma_{(l,k)} - 1]; p_{1,1(l,k)} = p_{1(l,k)} p_{T(l,k)} \end{aligned} \quad (10)$$

where $p_{T(l,k)} \triangleq 1 - \exp(-T_{(l,k)} / \lambda_{D(l,k)})$ denotes the probability of noise masked according to the auditory property. $T_{(l,k)}$ represents the masking threshold of auditory system [11]. $p_{T(l,k)}$ depends on the speech statistical model and details of its derivation is shown by Pu fanliang [12]. The gain of novel estimator under four-state

$$G = G_{H_{0,0}}^{p_{0,0}} G_{H_{0,1}}^{p_{0,1}} G_{H_{1,0}}^{p_{1,0}} G_{H_{1,1}}^{p_{1,1}} = G_{LS}^{p_1(1-p_T)} G_{\min}^{p_0(1-p_T)} \quad (11)$$

Noise spectrum estimation $\lambda_{D(l,k)}$ and a prior SNR $\xi_{(l,k)}$ are obtained by

$$\begin{aligned} \lambda_{D(l,k)} &= \left[\alpha_D \lambda_{D(l-1,k)} + (1-\alpha_D) |Y_{(l,k)}|^2 \right] p_{0(l,k)} (1-p_{T(l,k)}) \\ &\quad + |Y_{(l,k)}|^2 p_{0(l,k)} p_{T(l,k)} + \lambda_{D(l-1,k)} p_{1(l,k)} \\ &= \left[\alpha_D p_{0(l,k)} (1-p_{T(l,k)}) + p_{1(l,k)} \right] \lambda_{D(l-1,k)} \\ &\quad + \left[(1-\alpha_D) p_{0(l,k)} (1-p_{T(l,k)}) + p_{0(l,k)} p_{T(l,k)} \right] |Y_{(l,k)}|^2 \end{aligned} \quad (12)$$

$$\begin{aligned} \xi_{(l,k)} &= \left[\alpha A_{(l-1,k)}^2 / \lambda_{D(l-1,k)} + (1-\alpha) Q[\gamma_{(l,k)} - 1] \right] p_{1(l,k)} (1-p_{T(l,k)}) \\ &\quad + [\gamma_{(l,k)} - 1] p_{1(l,k)} p_{T(l,k)} + p_{0(l,k)} A_{(l-1,k)}^2 / \lambda_{D(l-1,k)} \\ &= \left[\alpha p_{1(l,k)} (1-p_{T(l,k)}) + p_{0(l,k)} \right] A_{(l-1,k)}^2 / \lambda_{D(l-1,k)} \\ &\quad + \left[(1-\alpha) p_{1(l,k)} (1-p_{T(l,k)}) + p_{1(l,k)} p_{T(l,k)} \right] Q[\gamma_{(l,k)} - 1] \end{aligned} \quad (13)$$

According to (12) and (13)

$$\tilde{\alpha}_{(l,k)} = \left[\alpha p_{1(l,k)} (1-p_{T(l,k)}) + p_{0(l,k)} \right] \quad (14)$$

$$\hat{\alpha}_{D(l,k)} = \left[\alpha_D p_{0(l,k)} (1-p_{T(l,k)}) + p_{1(l,k)} \right] \quad (15)$$

Smoothing parameters $\tilde{\alpha}_{(l,k)}$, $\hat{\alpha}_{D(l,k)}$ of novel algorithm are consistent with the rule employed by smoothing parameters to control the optimal tradeoff between musical noise level and audible distortion.

4. Experimental result

In this part, it can be testified that the new algorithm yields better result than OM-MMLS, Casual which new algorithm are based on. We compare them based on an objective improvement in the segmental SNR, a subjective study of speech spectrograms under several environments. Three different noise types, taken from Noisex92 database, are used in our evaluation: white Gaussian noise, m109 tank noise, and F16 cockpit noise. Each speech signal is degraded by the various noise types with segmental SNRs in the range [-5, 10] dB.

Table 1. Segmental SNR improvement for various noise types

Input SegSNR		5 dB	0 dB	-5 dB	-10 dB
White Noise	OM-MMLS	5.13	8.96	11.35	12.19
	Casual	7.21	9.83	12.52	13.61
	Novel	7.35	10.2	13.47	14.53
M109 Noise	OM-MMLS	4.25	7.29	8.89	9.69
	Casual	6.51	8.74	10.03	10.96
	Novel	7.1	9.33	10.91	12.08
F16 Noise	OM-MMLS	5	7.51	8.59	9.29
	Casual	5.61	8.87	9.93	11.28
	Novel	7.75	9.49	10.08	12.67

Table1 shows the average SegSNR improvement obtained for various noise types and at various noise levels. The novel algorithm's degree of SegSNR improvement is more obvious than other two algorithms, particularly for low input SNRs and nonstationary noise.

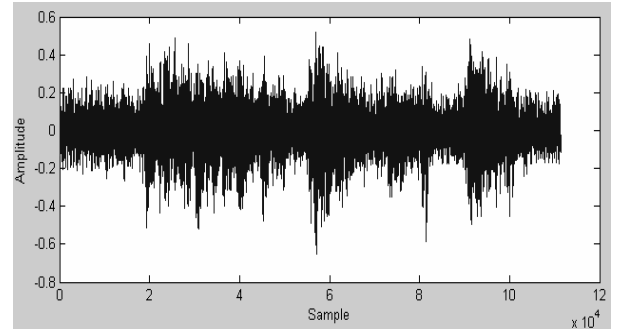


Figure 1.(a) Original white noisy speech of -5dB level

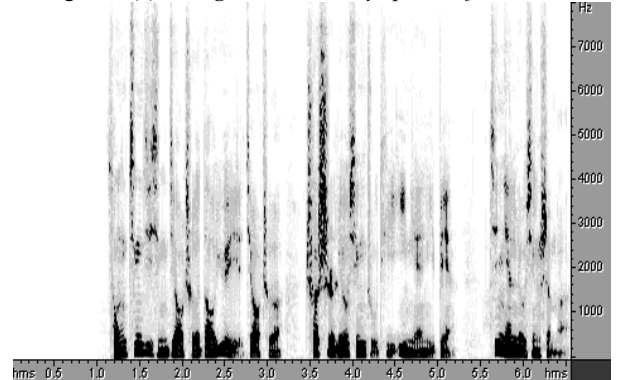


Figure 1.(b) The spectrogram of clean speech

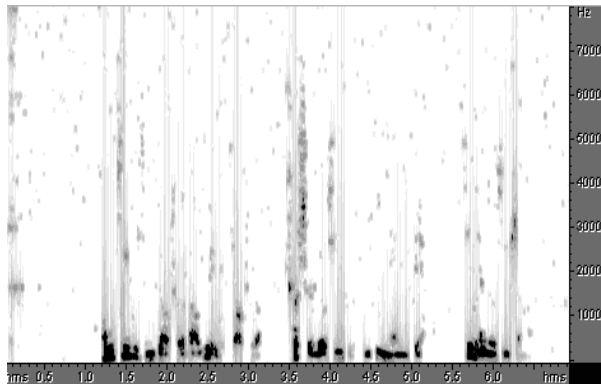


Figure 1.(c) *Speech enhanced by OM-MMLS*

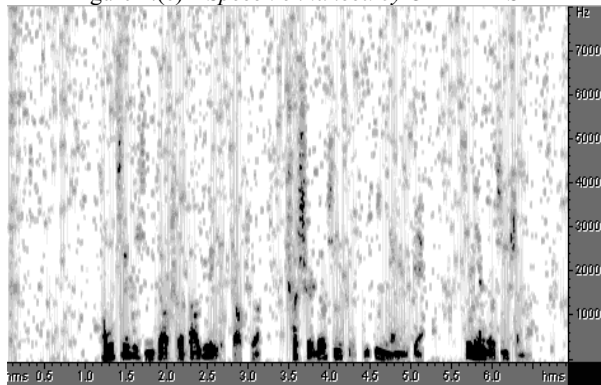


Figure 1.(d) *Speech enhanced by Casual*

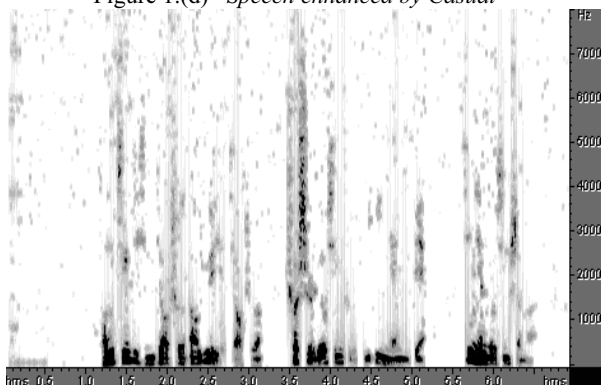


Figure 1.(e) *Speech enhanced by novel algorithm*

Figure 1: *Speech spectrograms of various enhanced speech*

Figure 1.(c) and Figure 1.(d) illustrate that OM-MMLS reduces most musical noise but distorts speech signal most, comparatively, Casual keeps more speech components but remains some perceptible musical noise. Figure 1.(e) demonstrates that enhanced speech of novel algorithm, compared with other two methods, remains most speech components while keeping the noise under the auditory masking threshold. It results in less audible distortion and less perceptible residual musical noise. Objective and subjective evaluation in various environment show that the proposed approach is advantageous. Excellent noise reduction can be achieved even in the most adverse noise conditions, while avoiding musical residual noise and the attenuation of weak speech components.

5. Conclusion

Smoothing parameters in recursive procedure of prior SNR and noise spectrum estimate are dominant factors to trade off musical noise level and speech distortion level. The novel algorithm presents new four-state speech uncertainty model

by incorporating masking property probability into speech uncertainty probability. On basis of that assumption, we modify log-spectral gain function, prior SNR, noise spectrum estimate and probability of each state under each speech state respectively. In addition, we employ probability of each speech state to adjust smoothing parameters more reasonably rather than set them as fixed value. It allows for an automatic adaptation in time and frequency of the parametric enhancement system, and finds optimal tradeoff among the remained perceptible noise level, musical noise level and speech distortion level based on a criterion correlated with perception.

6. References

- [1] Yariv. Ephraim, David. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32(6), pp. 1109-1121, December 1984.
- [2] Yariv. Ephraim, David. Malah, "Speech Enhancement Using a Minimum Mean Square Error Log-Spectral Amplitude Estimator", Transactions on Acoustics, Speech, and Signal Processing, vol.33 (2), pp. 443-445. April 1985.
- [3] Olivier. Cappe, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor", IEEE Transactions on Speech and Audio Processing, vol. 2(2), pp. 345-349, April 1994.
- [4] Israel.Cohen, "On Speech Enhancement under Signal Presence Uncertainty", Conf. Acoustics, Speech, and Signal Processing, Salt Lake City, UT, pp. 167-170, May 2001.
- [5] Israel. Cohen, "Optimal Speech Enhancement under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator", IEEE Signal Processing Letters, vol. 9(4), pp.113-116, April 2002.
- [6] Robert.J. Mcaclay, Marilyn.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28(2), pp. 137-145, April 1980.
- [7] Israel. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging", IEEE Transactions on Speech and Audio Processing, vol. 11(5), pp. 466-475, September 2003.
- [8] Israel. Cohen, "On the Decision-Directed Approach of Ephraim and Malah", ICASSP, pp. 293-296, 2004.
- [9] Israel. Cohen, "Speech Enhancement Using a Noncasual a Priori SNR Estimator", IEEE Signal Processing Letters, vol. 11(9), pp. 725-728, September 2004.
- [10] Israel. Cohen, "Relaxed Statistical Model for Speech Enhancement and a Prior SNR Estimation", IEEE Transactions on Speech and Audio Processing, vol. 13(5), pp. 870-881, September 2005.
- [11] Nathalie.Virag, "Signal Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", IEEE Transactions on Speech and Audio Processing, vol. 7(2), pp. 126-137, March 1999.
- [12] Zhengzhong Bian, Qijun Dai, Yanpu Chen, "Masked Noise Probability-based Speech Enhancement", Proceeding of the Second Joint EMBS/BMES Conference, Houston, TX, US, October23-26,2002.