Speech Enhancement for Nonstationary Noise Environments

Sandhya Hawaldar[1] and Manasi Dixit[2]

Department of Electronics, KIT's College of Engineering, Shivaji University, Kolhapur, 416234, India

santell2007@gmail.com

Abstract

In this paper, we present a simultaneous detection and estimation approach for speech enhancement in nonstationary noise environments. A detector for speech presence in the short-time Fourier transform domain is combined with an estimator, which jointly minimizes a cost function that takes into account both detection and estimation errors. Under speech-presence, the cost is proportional to a quadratic spectral amplitude error, while under speech-absence, the distortion depends on a certain attenuation factor. Experimental results demonstrate the advantage of using the proposed simultaneous detection and estimation approach which facilitate suppression of nonstationary noise with a controlled level of speech distortion.

Keywords

Estimation, Nonstationary noise, Spectral analysis, Speech enhancement, Decision rule.

1. Introduction

A practical speech enhancement system generally consists of two major components: the estimation of noise power spectrum, and the estimation of speech. The estimation of noise, when only one microphone source is provided, is based on the assumption of a slowly varying noise environment. In particular, the noise spectrum remains virtually stationary during speech activity. The estimation of speech is based on the assumed statistical model, distortion measure, and the estimated noise. A commonly used approach for estimating the noise power spectrum is to average the noisy signal over sections which do not contain speech. Existing algorithms often focus on estimating the spectral coefficients rather than detecting their existence. The spectral-subtraction algorithm [1] [2] contains an elementary detector for speech activity in the time–frequency domain, but it generates musical noise caused by falsely detecting noise peaks as bins that contain speech, which are randomly scattered in the STFT domain. Subspace approaches for speech enhancement [3] [4] decompose the vector of the noisy signal into a signal-plus-noise subspace and a noise subspace, and the speech spectral coefficients are estimated after removing the noise subspace. Accordingly, these algorithms are aimed at detecting the speech coefficients and subsequently estimating their values. McAulay and Malpass [5] were the first to propose a speech spectral estimator under a two-state model. They derived a maximum-likelihood (ML) estimator for the speech spectral amplitude under speech-presence uncertainty. Ephraim and Malah followed this approach of signal estimation under speech presence uncertainty and derived an estimator which minimizes the mean-square error (MSE) of the short-term spectral amplitude (STSA) [6]. In [7], speech presence probability is evaluated to improve the minimum MSE (MMSE) of the log-spectral amplitude (LSA) estimator, and in [8] a further improvement of the

DOI : 10.5121/sipij.2011.2411

MMSE-LSA estimator is achieved based on a two-state model. Under speech absence hypothesis, Cohen and Berdugo [8] considered a constant attenuation factor to enable a more natural residual noise, characterized by reduced musicality. Under slowly time-varying noise conditions, an estimator which minimizes theMSE of the STSA or the LSA under speech presence uncertainty may yield reasonable results [11].However, under quickly time-varying noise conditions, abrupt transients may not be sufficiently attenuated, since speech is falsely detected with some positive probability. Reliable detectors for speech activity and noise transients are necessary to further attenuate noise transients without much degrading the speech components. Despite the sparsity of speech coefficients in the time–frequency domain and the importance of signal detection for noise suppression performance, common speech enhancement algorithms deal with speech detection *independently* of speech estimation. Even when a voice activity detector is available in the STFT domain ,it is not straightforward to consider the detection errors when designing the optimal speech estimator.

High attenuation of speech spectral coefficients due to missed detection errors may significantly degrade speech quality and intelligibility, while falsely detecting noise transients as speech-contained bins, may produce annoying musical noise. In this paper, we present a simultaneous detection and estimation approach for speech enhancement in nonstationary noise environments. A detector for speech presence in the short-time Fourier transform domain is combined with an estimator, which jointly minimizes a cost function that takes into account both detection and estimation errors. Cost parameters control the tradeoff between speech distortion, caused by missed detection of speech components and residual musical noise resulting from false-detection. Under speech-presence, the cost is proportional to quadratic spectral amplitude (QSA) error [6], while under speech-absence, the distortion depends on a certain attenuation factor [2], [8], [9]. The noise spectrum is estimated by recursively averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability in subbands [13].

This paper is organized as follows. In Section 2 review of classical speech enhancement. In Section 3, proposed approach for speech enhancement. In Section 4 we compare the performance of the proposed approach to existing algorithms, both under stationary and nonstationary environments. In section 5, we conclude the advantages of simultaneous detection & estimation approach with modified speech absence estimate.

2. Classical Speech Enhancement

Let $\boldsymbol{x(n)}$ and $\boldsymbol{d(n)}$ denote speech and uncorrelated additive noise signals, and let $\boldsymbol{y(n)=x(n)+d(n)}$ be the observed signal.

Applying the STFT to the observed signal, we have,

$$\boldsymbol{Y_{lk} = X_{lk} + D_{lk}} \quad (1)$$

where $\boldsymbol{l=0,1,...}$ is the time frame index and $\boldsymbol{k=0,1,...,K-1}$ is the frequency-bin index. Let $\boldsymbol{H_1^{lk}}$ and $\boldsymbol{H_0^{lk}}$ denote, respectively, speech presence and absence hypotheses in the time–frequency bin(l, k) , i.e.,

$$\boldsymbol{H_1^{lk} : Y_{lk} = X_{lk} + D_{lk}}$$
$$\boldsymbol{H_0^{lk} : Y_{lk} = D_{lk}} \quad (2)$$

Assume that the noise expansion coefficients can be represented as the sum of two uncorrelated noise components: $\boldsymbol{D_{lk} = D_{lk}^s + D_{lk}^t}$ where $\boldsymbol{D_{lk}^s}$ denotes a quasi-stationary noise component, and $\boldsymbol{D_{lk}^t}$ denotes a highly nonstationary transient component. The transient components are

generally rare, but they may be of high energy and thus cause significant degradation to speech quality and intelligibility. But in many applications, a reliable indicator for the transient noise activity may be available in the system. For example, in an emergency vehicle (e.g., police or ambulance) the engine noise may be considered as quasi-stationary, but activating a siren results in a highly nonstationary noise which is perceptually very annoying. Given that a transient noise source is active, a detector for the transient noise in the STFT domain may be designed and its spectrum can be estimated based on training data.

The objective of a speech enhancement system is to reconstruct the spectral coefficients of the speech signal such that under speech-presence a certain distortion measure between the spectral coefficient and its estimate, $d_{ij}(X, \hat{X})$, is minimized, and under speech-absence a constant attenuation of the noisy coefficient would be desired to maintain a natural background noise [6], [9]. Most classical speech enhancement algorithms try to estimate the spectral coefficients rather than detecting their existence, or try to independently design detectors and estimators. The well-known spectral subtraction algorithm estimates the speech spectrum by subtracting the estimated noise spectrum from the noisy squared absolute coefficients [1], [2], and thresholding the result by some desired residual noise level. Thresholding the spectral coefficients is in fact a detection operation in the time–frequency domain, in the sense that speech coefficients are assumed to be absent in the low-energy time–frequency bins and present in noisy coefficients whose energy is above the threshold. McAulay and Malpass were the first to propose a two-state model for the speech signal in the time–frequency domain [5].The resulting estimator does not detect speech components, but rather, a soft-decision is performed to further attenuate the signal estimate by the *a posteriori* speech presence probability.

If an indicator for the presence of transient noise components is available in a highly nonstationary noise environment, then high-energy transients may be attenuated by using OM-LSA estimator [8] and setting the *a priori* speech presence probability to a sufficiently small value. Unfortunately, an estimation-only approach under signal presence uncertainty produces larger speech degradation, since the optimal estimate is attenuated by the *a posteriori* speech presence probability. On the other hand, increasing *a priori* speech presence probability prevents the estimator from sufficiently attenuating noise components. Integrating a jointly detector and estimator into the speech enhancement system may significantly improve the speech enhancement performance under nonstationary noise environments and allow further reduction of transient components without much degradation of the desired signal.

3. Proposed Approach for Speech Enhancement

Let $C_j\left(X, \hat{X}\right) \geq 0$ denote the cost of making a decision η_j and choosing an estimator $\hat{X}_j$ where X is the desired signal. Then, the Bayes risk of the two operations associated with simultaneous detection and estimation is defined by [11] and [12]

$$R = \sum_{j=0}^{1} \int_{\Omega_y} \int_{\Omega_x} C_j\left(X, \hat{X}\right) p\left(\eta_j \mid Y\right) p\left(Y \mid X\right) p\left(X\right) dX\, dY \tag{3}$$

where Ω_x and Ω_y are the spaces of the speech and noisy signals, respectively. The simultaneous detection and estimation approach is aimed at jointly minimizing the Bayes risk over both the decision rule and the corresponding signal estimate. Let $q \triangleq p(H_1)$ denote the *a priori* speech presence probability and let X_R and X_I denote the real and imaginary parts of the expansion coefficient X . Then, the *a priori* distribution of the speech expansion coefficient follows:

$$p(X) = q\,p(X \mid H_1) + (1-q)\,p(X \mid H_0) \tag{4}$$

where $p(X \mid H_0) = \delta(X)$ and $\delta(X) \equiv \delta(X_R, X_I)$ denotes the Dirac-delta function.
The cost function $C_j(X, \hat{X})$ may be defined differently whether H_1 or H_0 is true. Therefore, we let $C_{ij}(X, \hat{X}) \equiv C_j(X, \hat{X} \mid H_i)$ denote the cost which is conditioned on the true hypothesis.1 The cost function depends on both the true signal value and its estimate under the decision and therefore couples the operations of detection and estimation.

The cost function associated with the pair $\{H_j, \eta_j\}$ is generally defined by,

$$C_{ij}(X, \hat{X}) = b_{ij} d_{ij}(X, \hat{X}) \tag{5}$$

where $d_{ij}(X, \hat{X})$ is an appropriate distortion measure and the cost parameters b_{ij} control the tradeoff between the costs associated with the pairs $\{H_j, \eta_j\}$.That is, a high-valued b_{01} raises the cost of a false alarm, (i.e., decision of speech presence when speech is actually absent) which may result in residual musical noise. Similarly, b_{10} is associated with the cost of missed detection of a signal component, which may cause perceptual signal distortion. Under a correct classification, $b_{00} = b_{11} = 1$ normalized cost parameters are generally used. However $d_{ij}(.,.)$ is not necessarily zero since estimation errors are still possible even when there is no detection error. When speech is indeed absent, the distortion function is defined to allow some natural background noise level such that under H_0, the attenuation factor will be lower bounded by a constant gain floor $G_f \equiv 1$ as proposed in [2], [8], [9].

The distortion measure of the QSA cost function is defined by,

$$d_{ij}(X, \hat{X}) = \begin{cases} (|X| - |\hat{X}_j|)^2, i = 1 \\ (G_f |Y| - |\hat{X}_j|)^2, i = 0 \end{cases} \tag{6}$$

and is related to the STSA suppression rule of Ephraim and Malah [6].Assume that both X and D are statistically independent, zero-mean, complex-valued Gaussian random variables with variances λ_x and λ_d , respectively. Let $\xi \equiv \lambda_x / \lambda_d$, denote the *a priori* SNR under hypothesis H_1 , let $\gamma \equiv |Y|^2 / \lambda_d$,denote the *a posteriori* SNR and let $\upsilon \equiv \gamma\xi / (1+\xi)$. For evaluating the optimal detector and estimator under the QSA cost function we denote by $X \equiv ae^{j\alpha}$ and $Y \equiv Re^{j\theta}$ the clean and noisy spectral coefficients, respectively, where $a = |X|$ and $R = |Y|$.Accordingly, the pdf of the speech expansion coefficient under H_1 satisfies,

$$p(a, \alpha \mid H_1) = \frac{a}{\pi\lambda_x} \exp\left(-\frac{a^2}{\lambda_x}\right) \tag{7}$$

As proposed in [14], the optimal estimation under the decision $\eta_j, j \in \{0,1\}$

$$\hat{X}_j = \left[b_{1j}\Lambda(\xi,\gamma) G_{STSA}(\xi,\gamma) + b_{0j} G_f\right] \phi_j(\xi,\gamma)^{-1} Y \tag{8}$$

$$\equiv G_j(\xi,\gamma) Y$$

The optimal estimator under decision η_0 is modified with certain attenuation factor based on noise variance $\lambda_{d_{lk}}$ and Noisy speech power S_{lk} ,

$$\hat{X}_0 = G_0 \times Y \tag{9}$$

Where $G_0 = G_f \times (\lambda_{d_{lk}} / S_{lk} + e^{-10})^{1/2}$ and

To obtain S_{lk} recursive averaging is employed such that

$$S_{lk} = \xi_s S_{l-1,k} + (1-\xi_s)|Y_{lk}|^2$$

where ζs ($0 < \zeta s < 1$) is a smoothing parameter.
This modification reduces greatly the nonstationary noise from Noisy speech as it considers noisy speech power along with its variance.

The decision rule [14] under the QSA cost function is,

$$\Lambda(\xi,\gamma)\begin{cases} b_{10}G_0^2 - G_1^2 + \dfrac{\xi}{(1+\xi)\gamma}(1+\upsilon)(b_{10}-1) + \\ 2(G_1 - b_{10}G_0)G_{STSA} \end{cases} \begin{matrix} \eta_1 \\ > \\ < \\ \eta_0 \end{matrix} b_{01}(G_1 - G_f)^2 - (G_0 - G_f)^2 \tag{10}$$

Fig.1. shows a block diagram of the simultaneous detection and estimation system, the estimator is obtained by (8) and (9) and the interrelated decision rule (10) chooses the appropriate estimator for minimizing the combined Bayes risk.

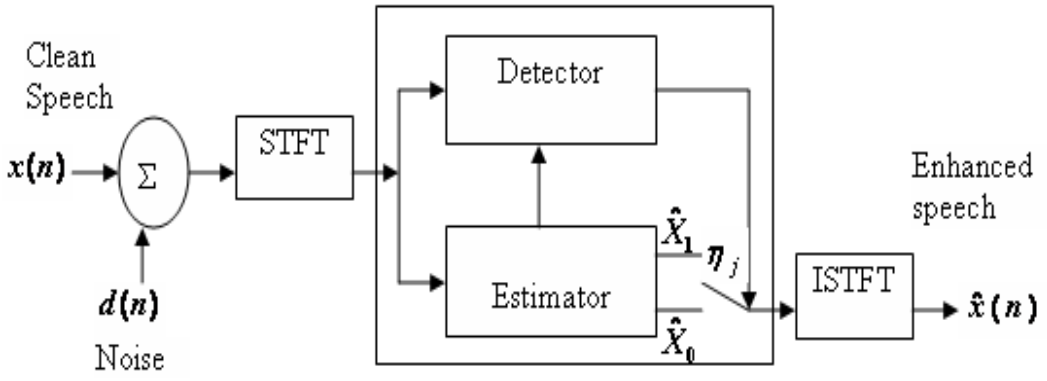

Fig.1. Simultaneous Detection and Estimation System.

4. Experimental Results

In our experimental study we consider the problem of hands free communication in an emergency vehicle and demonstrate the advantage of the simultaneous detection and estimation approach under stationary & nonstationary noise environments. Speech signals are recorded with sampling frequency at 8 kHz and degraded by different stationary & nonstationary additive noise. Nonstationary noise like siren noise is added with car noise for different levels of input SNR. The test signals include 12 speech utterances from 12 different speakers, half male and half female. The noisy signals are transformed into the STFT domain using half-overlapping Hamming windows of 32-ms length, and the background-noise spectrum is estimated by using the IMCRA algorithm[13] .The performance evaluation includes objective quality measure- SNR defined, in dB, a subjective study of spectrograms, and informal listening tests.

The proposed approach is compared with the OM-LSA algorithm [8]. The speech presence probability required for the OM-LSA estimator as well as for the simultaneous detection and estimation approach is estimated as proposed in [8]. For the OM-LSA algorithm, the decision-directed estimator with $\alpha = 0.92$ is implemented as specified in [8], and the gain floor

is $G_f = -20dB$. Fig. 2 shows waveforms and spectrograms of a clean signal, noisy signal, and enhanced signals for Speech degraded by car and siren noise with SNR of 5 dB. The speech enhanced by using the OM-LSA algorithm & the simultaneous detection and estimation approach are shown in Fig. 2.(c) and 2.(d), respectively. However, the simultaneous detection and estimation approach with modified speech absence estimate yields greater reduction of transient noise without affecting the quality of the enhanced speech signal.

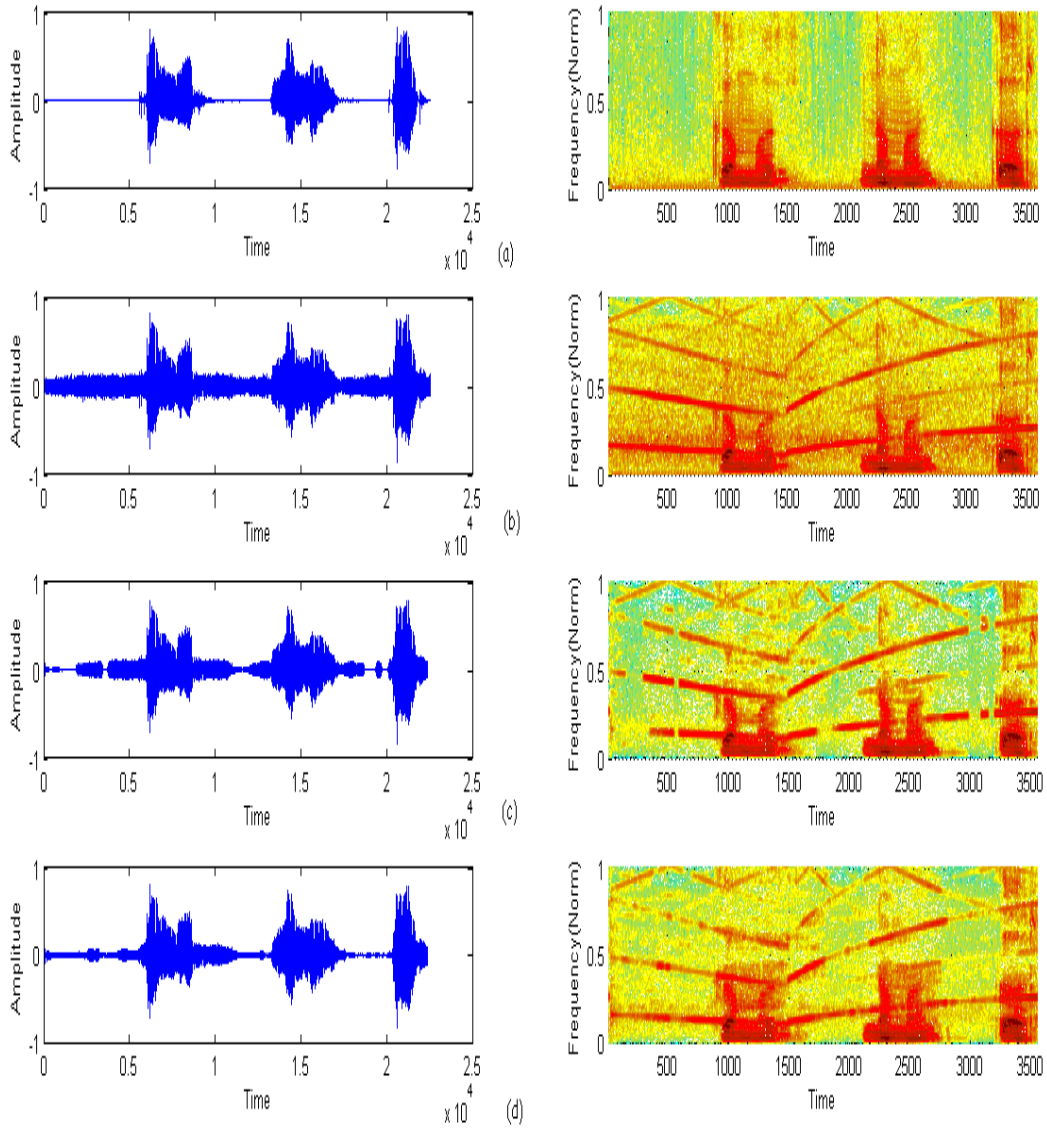

Fig. 2. Speech waveforms and spectrograms. (a) Clean speech signal: "kamal naman kar." in Marathi uttered by a male subject (b) Speech degraded by car noise and siren noise with SNR of 5dB. (c) Speech enhanced by using the OM-LSA estimator. (d) Speech enhanced by using the simultaneous detection and estimation approach with modified speech absence estimate using, b01 = b_{10} = 2.5 as proposed by authors.

Quality measures for the different input SNRs are shown in Table 1 & Table 2.The results from Table 1 demonstrate improved speech quality obtained by the simultaneous detection and estimation approach for stationary noise environments.

TABLE 1

Output SNR (in dB) By Using The OM-LSA Estimator & Simultaneous Detection And Estimation Approach for Different Stationary Noise Environments with Varying Input SNR Between 15dB to -5dB.

Noise	**Input SNR**	**OM-LSA Estimator**	**Proposed Simultaneous Detection & Estimation approach**
White Gaussian Noise	15	18.0290	21.1937
	10	15.0292	18.7496
	5	11.9830	9.7287
	0	8.5279	7.7648
	-5	5.9435	7.3687
Car	15	15.5912	20.8507
	10	12.9320	16.3415
	5	10.5166	13.8647
	0	8.3231	10.3499
	-5	6.0857	7.6548

The results from Table 2 demonstrate improved speech quality obtained by the simultaneous detection and estimation approach with modified speech absence estimate for nonstationary noise environments (car with siren noise).

TABLE 2

Output SNR (in dB) By Using The OM-LSA Estimator & Simultaneous Detection And Estimation Approach for Different Nonstationary Noise Environments with Varying Input SNR Between 15dB to -5dB.

Noise	**Input SNR**	**OM-LSA Estimator**	**Proposed Simultaneous Detection & Estimation approach**
Car(with siren noise)	15	16.1622	19.7296
	10	12.8231	16.0058
	5	12.3619	12.4352
	0	3.0348	4.9814
	-5	-3.9558	-2.3462
Train	15	17.6705	21.6112
	10	15.5157	17.0405
	5	14.9869	16.2158
	0	12.2862	13.4268
	-5	6.8280	7.7332

Subjective listening tests confirm that the speech quality improvement achieved by using the proposed method.

5. Conclusion

We have presented a single-channel speech enhancement approach in the time–frequency domain for nonstationary noise environments. A detector for the speech coefficients and a corresponding

estimator with modified speech absence estimate for their values is jointly designed to minimize a combined Bayes risk. In addition, cost parameters enable to control the tradeoff between speech quality, noise reduction, and residual musical noise. Experimental results show greater noise reduction with improved speech quality when compared with the OM-LSA suppression rules under stationary and nonstationary noise. It is demonstrated that under nonstationary noise environment, greater reduction of nonstationary noise components may be achieved by exploiting reliable information with simultaneous detection and estimation approach.

REFERENCES

[1] S. F. Boll, "Suppression of acousting noise in speech using spectral subtraction," IEEE Trans. Acoust.,Speech, Signal Process., vol.ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP'79, Apr. 1979, vol. 4, pp. 208–211.

[3] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," IEEE Signal Process. Lett.,vol. 10, no. 4, pp. 104–106, Apr. 2003.

[4] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," IEEE Trans. Speech Audio Process., vol. 11, no. 4, pp. 334–341, Jul. 2003.

[5] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft decision noise suppression filter," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[6] Y. Ephraim and 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.

[7] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments," in Proc. 24th IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP'99, Phoenix, AZ, Mar. 1999, pp. 789–792.

[8] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary environments,"Signal Process., vol. 81, pp. 2403–2418, Nov. 2001.

[9] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Process., vol. 2, no. 2, pp. 345–349, Apr. 1994.

[10] D. Middleton and F. Esposito, "Simultaneous optimum detection and estimation of signals in noise," IEEE Trans. Inf. Theory, vol. IT-14, no. 3, pp. 434–444, May 1968.

[11] Y. Ephraim and D. Malah , "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans.Acoust. , Speech, Signal Process., vol. 33, no. 2, pp. 443–445, Apr.1985.

[12] A. Fredriksen, D. Middleton, and D. Vandelinde , "Simultaneous signal detection and estimation under multiple hypotheses," IEEE Trans. Inf.Theory, vol. IT-18, no. 5, pp. 607–614, 1972.

[13] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," IEEE Trans. Speech Audio Process., vol. 11, no. 5, pp. 466–475, Sep. 2003.

[14] Ari Abramson and Israel Cohen ,"Simultaneous Detection and Estimation Approach for Speech Enhancement" IEEE Transactions On Audio, Speech, And Language Processing, Vol. 15, No. 8, Nov.2007