16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

SPEECH ENHANCEMENT USING INTRA-FRAME DEPENDENCY IN DCT DOMAIN

Achintya Kundu, Saikat Chatterjee and T.V. Sreenivas

Department of Electrical Communication Engineering,
Indian Institute of Science, Bangalore, India 560012.
phone: +91 80 2360 2167, Fax: +91 80 2360 0683.
email: {achintya, saikat, tvsree}@ece.iisc.ernet.in

ABSTRACT

In this paper, we present a new speech enhancement approach, that is based on exploiting the intra-frame dependency of discrete cosine transform (DCT) domain coefficients. It can be noted that the existing enhancement techniques treat the transform domain coefficients independently. Instead of this traditional approach of independently processing the scalars, we split the DCT domain noisy speech vector into sub-vectors and each sub-vector is enhanced independently. Through this sub-vector based approach, the higher dimensional enhancement advantage, viz. non-linear dependency, is exploited. In the developed method, each clean speech sub-vector is modeled using a Gaussian mixture (GM) density. We show that the proposed Gaussian mixture model (GMM) based DCT domain method, using sub-vector processing approach, provides better performance than the conventional approach of enhancing the transform domain scalar components independently. Performance improvement over the recently proposed GMM based time domain approach is also shown.

1. INTRODUCTION

Estimation of clean speech signal from noise corrupted speech is a challenging problem with applications in voice communication systems, automatic speech recognition systems, hearing aids, etc. Enhancement of noisy speech signal is generally carried out using statistical models of clean speech and noise. Existing approaches include spectral subtraction [1], Wiener filtering [2], Bayesian estimation approach in transform domain [3], hidden Markov model based methods [4], subspace based approach [5], etc. In this paper, we take the minimum mean square error (MMSE) estimation approach for speech enhancement (SE) in DCT domain.

For the traditional transform based SE methods [3], [6], [7], [8], the ubiquitous Gaussian density is used for modeling the probability densities of transform domain speech and noise coefficients. In the literature [9], it has been shown that the probability density function (PDF) of speech signal in signal/transform domain is non-Gaussian in nature. In [10], a DCT domain speech enhancement method is proposed based on modeling the PDF of clean speech DCT coefficients using Laplacian density. Among other non-Gaussian PDFs, Gamma distribution (family of super Gaussian densities) has been used in DFT/KLT domain [11]- [14]. In our recent work [15], we also have noted the importance of modeling the time domain speech coefficients using non-Gaussian PDF; we have modeled the joint PDF of time domain speech samples using GMM. It is mentioned that the GMM has been used earlier in speech enhancement to model the PDF of each short-time spectral component of speech [16],[17].

In transform domain, we note that the existing MMSE estimation based methods [3], [7], [8], [10], [13], enhance the transform domain coefficients of noisy speech individually, i.e., scalar processing is employed in the estimation process assuming the coefficients are independent. This approach will provide optimum performance if the respective joint PDFs of clean speech vector and noise vector can be effectively modeled using multivariate Gaussian densities (as the de-correlating transform makes the transform domain components independent). For signals with non-Gaussian PDF, there exists no linear transform which provides independent scalar components in transform domain. Thus, the transform-domain MMSE estimation method of enhancing scalar components independently leads to suboptimal performance for non-Gaussian PDF based signal, such as speech signal. To recover this performance loss, we investigate the approach of processing the sub-vectors in transform domain; the use of higher dimensional sub-vectors allows us to exploit the non-linear dependency which is otherwise not possible using scalar domain processing. In the developed method, the noisy speech signal vector is transformed using DCT and the DCT vector is split into sub-vectors; the sub-vectors are enhanced using the MMSE estimator. We have found that the new approach provides better performance than the conventional approach of enhancing the transform domain scalar components independently. Also, the new method has shown significant performance improvement over the recently proposed GMM based time domain method [15].

2. PROPOSED METHOD

We consider single-channel noisy speech signal as input to the speech enhancement system. Using additive model of speech signal degradation in noisy environment, input noisy speech signal can be written as

$$y(n) = x(n) + w(n),\ n = 0, 1, 2, \ldots, \qquad (1)$$

where $y(n)$, $x(n)$ and $w(n)$ are respectively nth sample of noisy speech signal, clean speech signal and additive noise. Speech enhancement system processes the sequence of noisy speech samples as overlapping frames, where each frame contains K consecutive samples and successive frames are shifted by R samples. We define tth noisy speech vector as $\mathbf{y}(t) = [y_t(0)\ y_t(1)\ \ldots\ y_t(K-1)]^T$, $t = 0, 1, \ldots$, where $y_t(n) = y(tR+n)$. Now, the noisy speech model of Eqn. (1) can be written in vector notation as $\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{w}(t)$, where $\mathbf{x}(t)$ and $\mathbf{w}(t)$ are $K \times 1$ vectors of clean speech and noise respectively corresponding to the noisy observation vector $\mathbf{y}(t)$. Denoting the $K \times K$ DCT matrix by $\mathbf{D}$, we define noisy speech vector, clean speech vector and

noise vector in DCT domain as $\mathbf{Y}(t) = \mathbf{D}\,\mathbf{y}(t)$, $\mathbf{X}(t) = \mathbf{D}\,\mathbf{x}(t)$, and $\mathbf{W}(t) = \mathbf{D}\,\mathbf{w}(t)$ respectively. Dropping the frame index t, we write the noisy speech model in DCT-domain as $\mathbf{Y} = \mathbf{X} + \mathbf{W}$, where $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{W}$ respectively denote $K \times 1$ random vectors of noisy speech, clean speech and additive noise in DCT domain. We view the speech enhancement problem as a statistical estimation problem, where we want to get an estimate of $\mathbf{X}$ from a given observation of $\mathbf{Y}$. We split the DCT-domain observation vector $\mathbf{Y}$ into S sub-vectors of dimension $L \times 1$ each. If ith sub-vector of noisy speech, clean speech and noise are respectively denoted by $\mathbf{Y}_i$, $\mathbf{X}_i$ and $\mathbf{W}_i$, then the observation model is given by

$$\mathbf{Y}_i = \mathbf{X}_i + \mathbf{W}_i, i = 1, 2, \ldots, S. \tag{2}$$

We assume that $\mathbf{X}_i$ and $\mathbf{X}_j$ are statistically independent for $i \neq j$. Similarly, $\mathbf{W}_i$ and $\mathbf{W}_j$ are also assumed to be independent for $i \neq j$. Again, the independence assumption between clean speech signal and corrupting noise says that $\mathbf{X}_i$ and $\mathbf{W}_j$ are statistically independent for any i, j. Therefore, $\mathbf{X}_i$ is independent of $\mathbf{Y}_j$ for $i \neq j$ and the MMSE estimator of $\mathbf{X}_i$ becomes

$$\widehat{\mathbf{X}}_i = \mathbf{E}\{\mathbf{X}_i|\mathbf{Y}\} = \mathbf{E}\{\mathbf{X}_i|\mathbf{Y}_i\}, \tag{3}$$

where $\mathbf{E}$ is the statistical expectation operator. We denote the PDFs of $\mathbf{X}_i$ and $\mathbf{W}_i$ by $f_{\mathbf{X}_i}(\mathbf{x})$ and $f_{\mathbf{W}_i}(\mathbf{w})$ respectively. In this method, we use Gaussian mixture density to model the PDF of each clean speech sub-vector $\mathbf{X}_i$ and a Gaussian density to model the PDF of each noise sub-vector $\mathbf{W}_i$ as shown below.

$$f_{\mathbf{X}_i}(\mathbf{x}_i) = \sum_{m=1}^{M_i} \alpha_{m,i}\, \mathbf{N}\left(\mathbf{x}_i; \mu^m_{\mathbf{X},i}, C^m_{\mathbf{XX},i}\right), \tag{4}$$

$$f_{\mathbf{W}_i}(\mathbf{w}_i) = \mathbf{N}\left(\mathbf{w}_i; \mu_{\mathbf{W},i}, C_{\mathbf{W},i}\right), 1 \leq i \leq S, \tag{5}$$

where $\alpha_{m,i}$, $\mu^m_{\mathbf{X},i}$ and $C^m_{\mathbf{XX},i}$ respectively denote prior probability, mean vector, covariance matrix of mth Gaussian component and M_i is the number of mixture component in the GMM of ith clean speech sub-vector. We mention that $\sum_{m=1}^{M_i} \alpha_{m,i} = 1$, $1 \leq i \leq S$. Mean vector and covariance matrix of ith noise sub-vector are denoted as $\mu_{\mathbf{W},i}$ and $C_{\mathbf{W},i}$ respectively. Now, for evaluating the expectation of Eqn. (3), we use our previously derived result [15] on GMM based MMSE estimation. By using the theorem of [15], we get

$$\widehat{\mathbf{X}}_i = \mathbf{E}\{\mathbf{X}_i|\mathbf{Y}_i\} = \sum_{m=1}^{M_i} \beta_{m,i}(\mathbf{Y}_i)\; \mu^m_{\mathbf{X}_i|\mathbf{Y}_i}(\mathbf{Y}_i), \tag{6}$$

where $\mu^m_{\mathbf{X}_i|\mathbf{Y}_i}(\mathbf{Y}_i)$ and $\beta_{m,i}(\mathbf{Y}_i)$ are respectively defined in Eqn. (7) and Eqn. (8). By concatenating the estimated sub-vectors, we get estimate of $\mathbf{X}$ as $\widehat{\mathbf{X}} = \left[\widehat{\mathbf{X}}_1^T\, \widehat{\mathbf{X}}_2^T \cdots \widehat{\mathbf{X}}_S^T\right]^T$. The enhanced speech vector is found by using inverse DCT as $\widehat{\mathbf{x}} = \mathbf{D}^{-1}\widehat{\mathbf{X}}$. Finally, enhanced speech frames are overlap-added to generate the output speech signal.

2.1 Motivation for split vector approach

In our recent paper [15], we have proposed a GMM based estimation frame-work for speech enhancement. There we have modeled the joint PDF of time-domain clean speech samples corresponding to a speech frame or vector. This allows us to exploit the dependency of speech samples in time-domain. Experimental result showed that using large number (1024) of mixture components in the GMM, speech enhancement performance increases with frame size (K) upto a certain value (5 msec) and then saturates for higher dimension due to the algorithmic problem in training GMM for high dimensional speech vector. But, we know that speech signal is nearly stationary over 20-40 msec segments. To exploit the statistical dependency of speech samples over such segments, larger frame size is generally preferred for improving the speech enhancement performance. To achieve the high dimensional advantage, we choose frame size of 32 msec ($K = 256$ samples for 8 kHz sampled speech). We avoid the problem of evaluating the GMM parameters in higher dimension by using the de-correlating transform (DCT). Thus, we split the larger dimensional transformed vector into smaller dimensional ($L \ll K$) sub-vectors and a separate GMM is trained for each sub-vector. The use of DCT helps in exploiting the correlation among time-domain speech samples over large frame, but the non-linear dependency among DCT-coefficients is exploited through the use of joint PDF for the coefficients of each sub-vector. Thus, the problem of training the higher dimensional GMM is avoided without sacrificing the performance.

3. EXPERIMENTS AND RESULTS

3.1 Experimental setup

The speech data used in the experiments are taken from the TIMIT database. The speech signal is first low pass filtered (3.4 kHz cut-off frequency) and then down-sampled to 8 kHz. We have used about 40 minutes of speech data for training and a separate 3 minutes of speech data (6 male and 6 female speakers speaking 5 sentences each) for testing. The training data is used for estimating the GMM parameters employing expectation-maximization (EM) algorithm. The test speech is generated by adding noise to the clean speech signal at the required level. We have considered different types of noise taken from NOISEX-92 database. In all our experiments, we have assumed the noise to be stationary and noise statistics are estimated only once from the initial 320 msec segment (containing only noise) of the test speech. We have chosen frame size (K) of 256 samples. To measure the speech enhancement performance, we use the following widely used objective measures : signal-to-noise ration (SNR), average segmental SNR (avg. seg. SNR) and perceptual evaluation of speech quality (PESQ) measure. Informal listening tests are conducted to verify the subjective quality of enhanced speech.

3.2 Performance of split-vector approach

We examine the effect of splitting the DCT domain vector on speech enhancement performance. Using the proposed DCT-domain GMM based method, we evaluate the speech enhancement performance in terms of SNR, average segmental SNR and PESQ score for various sub-vector dimensions (L) under different noisy conditions as shown in Fig. 1. We mention that we have used $M = 2L$ number of mixture components in the GMM to model the PDF of a L dimensional clean speech sub-vector. We know that by exploiting the dependency among the coefficients better performance can be

$$\mu^m_{\mathbf{x}_i|\mathbf{Y}_i}(\mathbf{Y}_i) = \mu^m_{\mathbf{x},i} + C^m_{\mathbf{xx},i}\left[C^m_{\mathbf{xx},i} + C_{\mathbf{w},i}\right]^{-1}\left(\mathbf{Y}_i - \left[\mu^m_{\mathbf{x},i} + \mu_{\mathbf{w},i}\right]\right),\ 1 \le m \le M_i, 1 \le i \le S. \tag{7}$$

$$\beta_{m,i}(\mathbf{Y}_i) = \frac{\alpha_{m,i}\,\mathbf{N}\left(\mathbf{Y}_i; \mu^m_{\mathbf{x},i} + \mu_{\mathbf{w},i}, C^m_{\mathbf{xx},i} + C_{\mathbf{w},i}\right)}{\sum_{j=1}^{M_i}\alpha_{j,i}\,\mathbf{N}\left(\mathbf{Y}_i; \mu^j_{\mathbf{x},i} + \mu_{\mathbf{w},i}, C^j_{\mathbf{x},i} + C_{\mathbf{w},i}\right)},\ 1 \le m \le M_i, 1 \le i \le S. \tag{8}$$

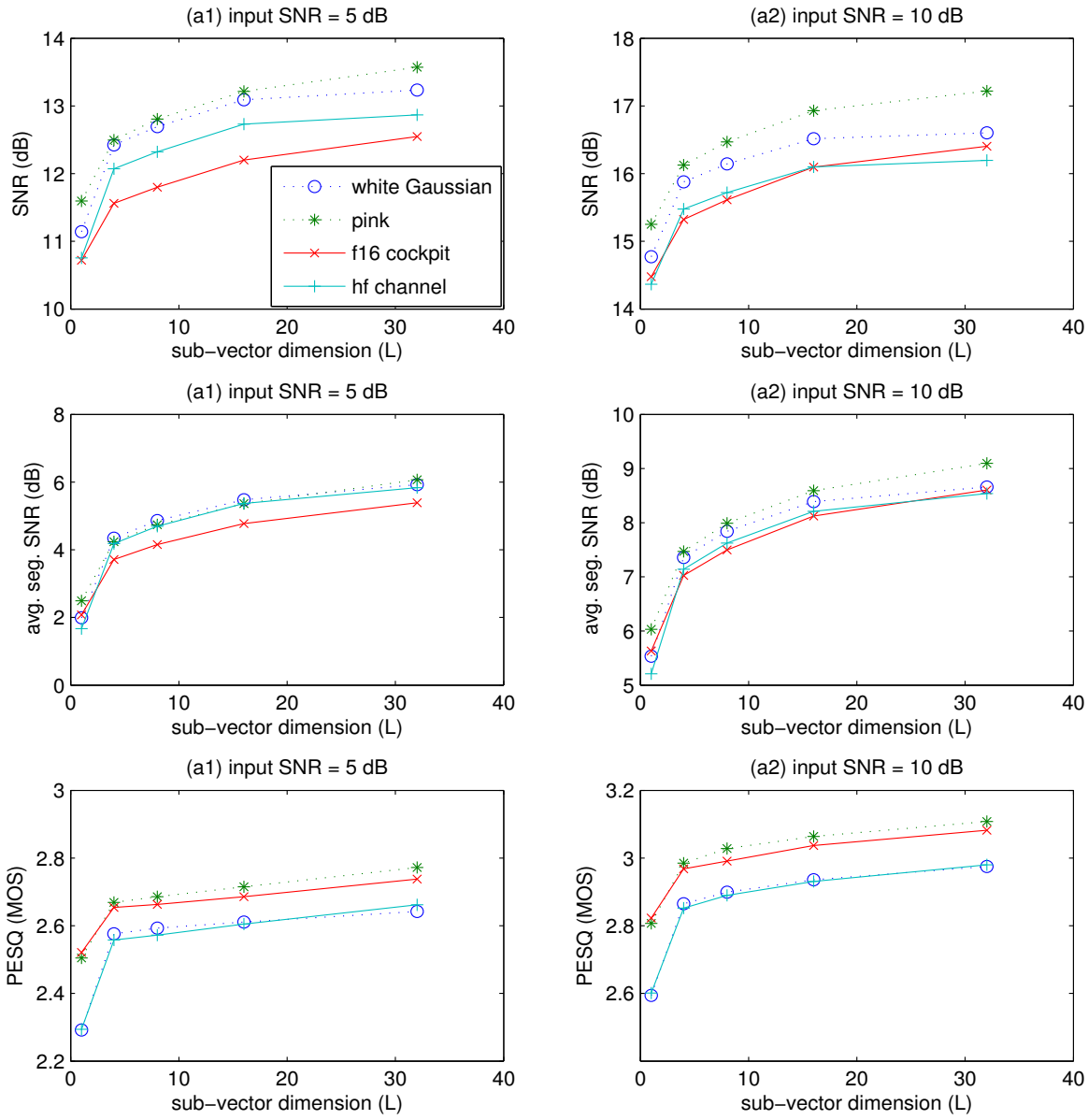

Figure 1: Speech enhancement performance in terms of SNR [(a1), (a2)], average segmental SNR [(b1), (b2)] and PESQ score [(c1), (c2)] at different sub-vector dimension (L = 1, 4, 8, 16, 32) for various types of noise.

obtained as illustrated in Fig. 1. We note that the performance improves monotonically as the sub-vectors' dimensions are increased. This observation clearly shows the advantage of sub-vector based processing ($L > 1$) over scalar processing ($L = 1$ case) in DCT domain. Another interesting observation from Fig. 1 is that the performance saturates as L is increased beyond a certain value (about $L = 10$). It is mentioned that the best performance can be achieved if the full vector is processed (i.e. without splitting). But, the saturation in performance with L suggests that the approach of splitting does not result in sacrificing the performance. Rather, this approach allows us to handle larger dimensional speech frames to exploit the statistical dependencies among the speech coefficients.

3.3 Performance comparison

We compare the performances of the developed GMM based DCT-domain method (using $L = 32$) with the recently proposed GMM based time-domain method [15]. We also include the performances of another DCT-domain method [7], which is based on modeling the PDF of DCT domain coefficients using Gaussian density. In Table 1, we present the speech enhancement performance of the three methods under various noisy conditions. Compared to the existing methods, the new method is shown to provide significant improvement in performance. The improvement in performance for the proposed DCT-domain method over the time-domain method is due to its ability to exploit the statistical dependency of speech samples over a much larger frame. Note that, for the time-domain method, the frame size of 40 samples is found to be optimum [15]). On the other hand, performance improves over the method of [7], because of using non-Gaussian PDF (Gaussian mixture) in DCT-domain rather than Gaussian density and applying the MMSE estimation on sub-vectors instead of individual coefficients. We also compare the proposed method with the DFT-domain method of Ephraim and Malah [3]. From Fig. 2, we note that the proposed method provides better speech enhancement performance in terms of all the three objective measures and for all types of noise. Informal listening tests also confirm about the superiority of the developed method. Thus, GMM based sub-vector processing in DCT-domain achieves higher SE performance than the conventional approaches.

4. CONCLUSIONS

A novel speech enhancement method exploiting the intra-frame dependency of clean speech DCT coefficients is proposed. The developed method nicely overcome the shortcomings of the recently proposed GMM based time-domain speech enhancement approach. Significant performance improvement in the proposed method over other speech enhancement methods is noted through the objective measurements and informal listening tests. Thus, DCT domain sub-vector enhancement approach using GMM can be regarded as an effective method for enhancing noisy speech signal.

REFERENCES

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

[2] P.C. Loizou, "Speech enhancement: Theory and practice," CRC Press, 2007.

[3] 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, pp. 1109-1121, Dec. 1984.

[4] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725-735, April 1992.

[5] Y. Ephraim and H.L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251-266, July 1995.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, pp. 443-445, April 1985.

[7] I.Y. Soon, S.N. Koh and C.K. Yeo, "Noisy speech enhancement using discrete cosine transform", *Speech Commun.*, vol. 24, No. 3, pp. 249-257, June 1998.

[8] J-H Chang and N.S. Kim, "Speech enhancement using warped discrete cosine transform," *Proc. IEEE Speech Coding Workshop*, pp. 175-177, Oct. 2002.

[9] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204-207, July 2003.

[10] S. Gazor and W. Zhang, "Speech enhancement employing Laplacian-Gaussian mixture," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 896-904, Sept. 2005.

[11] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 253-256, 2002.

[12] C. Breithaupt and R. Martin, "MMSE estimation of magnitude-squared DFT coefficients with superGaussian priors," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 896-899, April 2003.

[13] J.S. Erkelens, R.C. Hendriks, R. Heusdens and J. Jensen, "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, August 2007.

[14] J. Jensen and R. Heusdens, "Improved subspace-based single-channel speech enhancement using generalized super-Gaussian priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 862-872, March 2007.

[15] A. Kundu, S. Chatterjee, A.S. Murthy and T.V. Sreenivas, "GMM based Bayesian approach to speech enhancement in signal/transform domain," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 4893-4896, March 2008.

[16] I. Potamitis, N. Fakotakis and G. Kokkinakis, "A trainable speech enhancement technique based on mixture models for speech and noise," *Proc. EUROSPEECH*, pp. 573-576, Sept. 2003.

[17] G.H. Ding, X. Wang, Y. Cao, F. Ding and Y. Tang, "Speech enhancement based on speech spectral complex Gaussian mixture model," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 165-168, March 2005.

Table 1: Performances of recently proposed GMM based time-domain method ($Output_1$) of [15], new GMM based DCT-domain method ($Output_2$) and traditional Gaussian based DCT-domain method ($Output_3$) of [7]

Noise Type	SNR (dB)				Avg. Seg. SNR (dB)				PESQ (MOS)			
	Input	$Output_1$	$Output_2$	$Output_3$	Input	$Output_1$	$Output_2$	$Output_3$	Input	$Output_1$	$Output_2$	$Output_3$
white	0	8.92	10.05	8.13	-9.50	2.34	3.37	2.20	1.58	2.10	2.32	2.12
Gaussian	5	11.99	13.23	11.14	-4.50	4.68	5.92	4.68	1.84	2.33	2.64	2.50
noise	10	15.18	16.60	14.38	0.49	6.75	8.66	7.27	2.15	2.54	2.97	2.83
	0	8.75	10.05	7.96	-9.31	1.75	3.20	2.09	1.78	2.08	2.42	2.32
pink	5	12.01	13.57	11.14	-4.31	4.08	6.06	4.55	2.11	2.35	2.77	2.67
noise	10	15.43	17.22	14.67	0.68	6.83	9.09	7.28	2.45	2.61	3.10	3.00
f16	0	7.48	8.76	7.14	-9.20	0.79	2.28	1.32	1.86	2.11	2.38	2.22
cockpit	5	11.01	12.55	10.47	-4.20	3.42	5.39	3.99	2.17	2.38	2.73	2.61
noise	10	14.71	16.40	14.09	0.80	6.48	8.60	6.86	2.50	2.64	3.08	2.94
high freq.	0	8.61	9.74	7.59	-9.47	1.71	3.32	1.87	1.64	2.18	2.34	2.04
channel	5	11.77	12.87	10.69	-4.47	4.38	5.83	4.42	1.89	2.40	2.66	2.44
noise	10	15.06	16.19	13.95	0.52	6.90	8.54	7.05	2.18	2.61	2.98	2.80
m109	0	13.46	15.03	12.35	-8.73	5.78	7.56	5.22	2.37	2.53	3.00	2.89
tank	5	16.32	18.58	15.54	-3.73	8.14	10.63	8.01	2.70	2.81	3.34	3.22
noise	10	19.37	22.19	19.01	1.26	10.87	13.87	11.08	3.03	3.11	3.65	3.51
speech	0	6.08	6.21	5.11	-8.83	0.12	-0.03	-0.39	1.94	2.01	2.01	1.94
babble	5	9.70	10.38	8.70	-3.83	2.74	3.54	2.43	2.27	2.34	2.41	2.34
noise	10	13.50	14.70	12.53	1.16	5.91	7.26	5.52	2.59	2.65	2.80	2.73

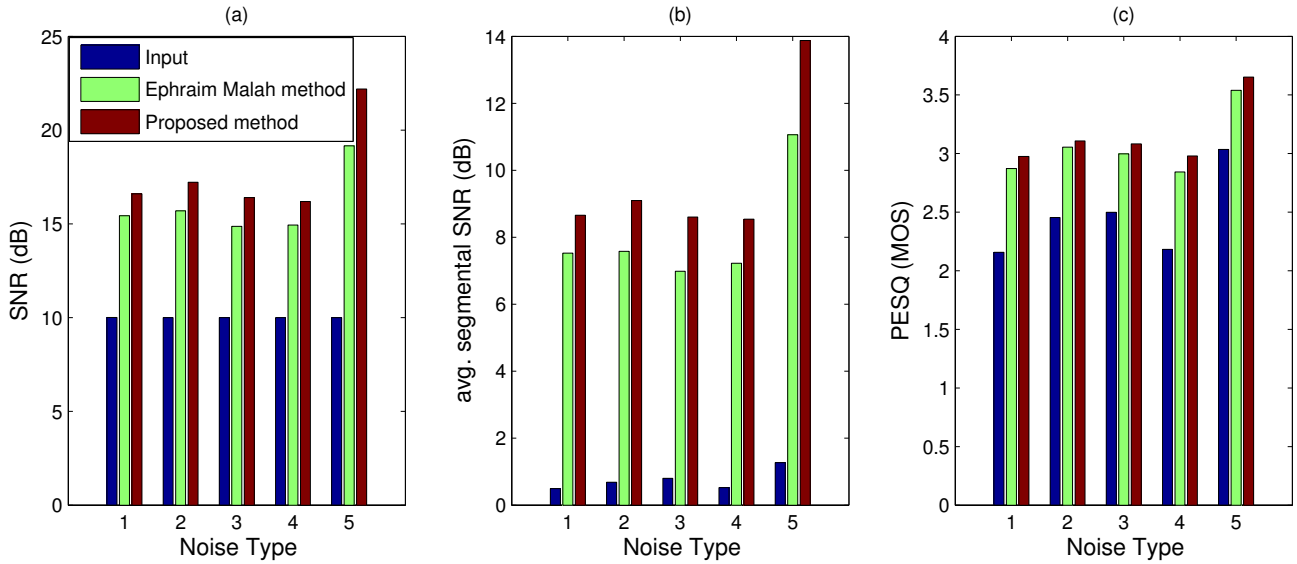

Figure 2: Speech enhancement performances of the Ephraim Malah method [3] and the proposed method for various types of noise: (1) white Gaussian, (2) pink, (3) f16 cockpit, (4) high freq. channel and (5) m109 tank noise.