TWO-STEP TECHNIQUE BASED LAPLACIAN FACTOR ESTIMATION FOR SPEECH ENHANCEMENT

SHIFENG OU, XIANYUN WANG, XIAOJUN ZHANG, YING GAO

Institute of Science and Technology for Opto-electronic Information, Yantai University, Yantai 264005, Shandong, China

ABSTRACT

The estimation of Laplacian factor is a crucial part of speech enhancement algorithms using Laplacian model priori. Classical methods for the estimation of this parameter suffer from the residual noise or time delay bias. In this paper, a novel algorithm called two-step technique for the estimation of Laplacian factor is proposed in discrete cosine transform domain. Based on the estimation theory of minimum mean square error (MMSE), the proposed algorithm computes directly the amplitude of the clean speech component to refine the estimated Laplacian factor in the original method, and thus the accurate result can be attained. Simulation results under different kinds of noisy conditions and levels demonstrate that the proposed algorithm possesses improved performance than that of the original method.

Keywords: *Speech Enhancement, Two-Step, Discrete Cosine Transforms, Noise Reduction, Laplcian Model*

1. INTRODUCTION

The purpose of speech enhancement is to improve the naturalness and perceptual quality for the proposed speech signal in order to reduce the fatigue of human listeners. It also aims at achieving a better intelligibility of the proposed speech for listeners or to increase the accuracy of a speech recognition system operating in a noisy environment. Many effective algorithms, such as spectral subtraction, hard or soft threshold, and minimum mean square error (MMSE) estimation and Wiener filtering, have been proposed, implemented, and reported in the last three decades [1-4]. However, the problem of speech enhancement remains open.

Most of the researches on noisy speech enhancement work in the Ddiscrete Fourier transform (DFT) to make it easier to remove noise embedded in the noisy speech signal [5]. This is often done as it is easier to separate the speech energy and the noise energy in the transform domain. Recently, it is reported that the discrete cosine transform (DCT) outperforms the DFT in terms of speech energy compaction, and the speech enhancement system using DCT has better noise reduction performance than that of the DFT based system [6]. Classical speech enhancement methods employing DCT are obtained based on the assumption that both the noise and the original speech signal amplitudes can be modeled by zero mean Gaussian distributed random variables in the transform domain. This assumption is supported by the central limit theorem as each transform coefficient is just a weighted sum of the speech samples. However, recent research has shown that in discrete cosine transform domain the Laplacian distribution is more suitable than the conventional Gaussian distribution for DCT coefficients of clean speech [7]. Based on this research, [8] proposed MMSE and Maximum Likelihood (ML) estimator for speech enhancement employing the Laplacian-Gaussian mixture model for noisy speech signal, which was shown to result in better performance for noise reduction compared to other methods under Gaussian model. However, the estimation of the Laplacian factor of the speech is derived using the noisy speech signal instead of the clean speech, so the resulting Laplacian factor estimation is not accurate because of the interference of noise energy. To further improve the performance, this paper proposes a new approach for estimating the Laplacian factor called two-step estimation technique. Based on the estimation theory of MMSE, the presented algorithm computes directly the amplitude of the clean speech component to refine the estimated Laplacian factor in [8], and thus the accurate result can be attained. The simulation experiment results validate that the proposed approach can availably improve the performance of a speech enhancement algorithm in DCT domain. The remained of this paper is organized as follows: Section 2 gives the speech enhancement method using Laplacian-Gaussian densities and a two-step technique for Laplacian factor estimation are proposed in section 3. In section 4, some tests are conducted to evaluate the performance of the algorithms, and finally in section 5, some concluding remarks are drawn.

2. SPEECH ENHANCEMENT EMPLOYING LAPLACIAN-GAUSSIAN PRIOR

N point DCT components $C(k)$, $0 \le k \le N-1$ of a length-N input sequence $f(n)$, $0 \le k \le N-1$ is defined by

$$C(k) = \frac{1}{\sqrt{N}} \mu_k \sum_{n=0}^{N-1} f(n) \cos\left(\frac{\pi}{2N}(2n+1)k\right) \quad (1)$$

where $\mu_0 = 1$, $\mu_k = \sqrt{2}$, for $1 \le k \le N-1$, and the inverse transformation is given by

$$f(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \mu_k C(k) \cos\left(\frac{\pi}{2N}(2n+1)k\right) \quad (2)$$

Here, the noise signal is assumed additive and uncorrelated with the clean speech signal, and N-dimensional vector of noisy speech at frame time m can be denoted

$$\boldsymbol{y}(m) = \boldsymbol{x}(m) + \boldsymbol{v}(m) \quad (3)$$

where $\boldsymbol{x}(m)$ and $\boldsymbol{n}(m)$ represent the N-dimensional clean speech vector and noise vector. Applying the DCT in (1) to the observed signal $\boldsymbol{y}(m)$, we have in the DCT domain

$$Y(k,m) = X(k,m) + N(k,m) \quad (4)$$

where $X(k,m)$, $Y(k,m)$ and $N(k,m)$ denote the DCT transformed components of the clean speech, noisy speech and noise signals respectively, K is the total number of frequency components, k and m represent the frequency and frame index. With the assumption that different DCT components along k and m are statistically independent, the MMSE estimated component $\hat{X}(k,m)$ can be obtained from the noisy component $Y(k,m)$ as follows (for the simplicity of notation, the index k and m are dropped in this paper).

$$\hat{X} = E\{X|Y\} = \frac{\int_{-\infty}^{\infty} X \cdot p\{Y|X\} p\{X\} \mathrm{d}X}{\int_{-\infty}^{\infty} p\{Y|X\} p\{X\} \mathrm{d}X} \quad (5)$$

where $p\{X\}$ and $p\{Y|X\}$ are the PDF of X and conditional distribution of Y given X, respectively, and $E(\cdot)$ denote the expectation operator.

We assume that the components of clean speech and noise can be modeled by a Laplacian and Gaussian distribution, respectively, the PDF of X and N are given by [8]

$$p\{X\} = \frac{1}{2a} \exp\left(-\frac{|X|}{a}\right) \quad (6)$$

$$p\{N\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{X^2}{2\sigma^2}\right) \quad (7)$$

where $a = E\{|X|\}$ is the Laplacian factor which represents the mean absolute value of clean speech component, and $\sigma^2 = E\{|N|^2\}$ is the variance of the noise component. Then the conditional distribution of Y given X is obtained as

$$p\{Y|X\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|Y-X|^2}{2\sigma^2}\right) \quad (8)$$

Substituting (6) and (8) into (5), the MMSE estimator for speech component is given by

$$\hat{X} = a e^{\frac{\psi}{2}} \left[\frac{(\psi+\xi)e^{\xi}\operatorname{erfc}(A) - (\psi-\xi)e^{-\xi}\operatorname{erfc}(B)}{e^{\xi}\operatorname{erfc}(A) + e^{-\xi}\operatorname{erfc}(B)}\right] \quad (9)$$

where $\xi = \frac{v}{a}$, $\psi = \frac{\lambda_u}{a^2}$, $A = \frac{\psi+\xi}{\sqrt{2\psi}}$, $B = \frac{\psi-\xi}{\sqrt{2\psi}}$, and the function $\operatorname{erfc}(t) = \frac{2}{\sqrt{\pi}} \int_t^{\infty} \exp(-x^2)\, dx$ is the complementary error function.

At the end, inverse DCT is performed on the estimated speech components and an estimation of the clean speech signal is synthesized.

3. TWO-STEP ESTIMATION TECHNIQUE

To use the equations (9) for noisy speech enhancement, the parameters σ^2 and a should to be estimated firstly. The value of σ^2 can is assumed to be known as it can be easily obtained using noise samples collected during speech absent frames, and the Laplacian factor $a(m)$ for clean speech component of the m-th frame can be estimated as the following equation

$$\hat{a}(m) = \beta\, \hat{a}(m-1) - (1-\beta)|X(m)| \quad (10)$$

where $\hat{a}(m-1)$ represents the estimated value of a resulting from the processing of the previous frame, and the weighting factor β is a time constant. Note that there is no access to the clean speech component $X(m)$, paper [8] proposed to use $Y(m)$ as an approximation for $X(m)$, i.e.

$$\hat{a}(m) = \beta\, \hat{a}(m-1) + (1-\beta)\left|Y(m)\right| \tag{11}$$

As the noisy speech component $Y(m)$ is used for estimating the Laplacian factor, the resulting $\hat{a}(m)$ in above equation will be generally bigger than its best value estimated from $X(m)$ because of the interference of the noise energy. Some research has investigated the effect of the estimation error of the Laplacian factor on enhanced speech, and found that the speech enhancement system using the estimated Laplacian factor described in (11) will generally leave more noise in the enhanced speech signals. To remove the effect of noise energy, [9] proposed an indirect estimation approach for improving the Laplacian factor estimation by using the generalized Gaussian density function described as follows

$$p\{z\} = \left\{\frac{r\eta(r,\sigma)}{2\Gamma(1/r)}\right\}\exp\left(-\left[\eta(r,\sigma)|z|\right]^{r}\right) \tag{12}$$

where

$$\eta(r,\sigma) = \sigma^{-1}\left[\frac{\Gamma(3/r)}{\Gamma(1/r)}\right]^{1/2} \tag{13}$$

z is a random variable, $\Gamma(\cdot)$ denotes the gamma function and σ is a positive real valued parameter, r is a shape parameter describing the exponential rate of decay. For the special cases r=1 the generalized Gaussian function $p\{z\}$ becomes a Laplacian PDF. Based on the relationship between the Laplacian factor and the variance of clean speech components under the Laplacian assumption model *i.e.* r =1, the estimation for Laplacian factor can be obtained by the following steps

$$\hat{\lambda}_X(m) = \eta\, \hat{X}^2(m-1) + (1-\eta)\max\left\{Y^2(m)-\sigma^2, 0\right\} \tag{14}$$

$$\hat{a}(m) = \frac{\sqrt{2}}{2}\sqrt{\hat{\lambda}_X(m)} \tag{15}$$

where $\hat{\lambda}_X$ is the estimated variance of clean speech component, $\hat{X}(m-1)$ is the estimated variance of the previous frame. It is well known that the decision-directed approach used in (14) has a serious drawback that the estimated result follows the shape of the real value with a simple delay of one short time frame, and thus the following estimation in (15) can not get an accurate result. To remove this bias, in this paper, we propose a new estimation approach for Laplacian factor called two-step estimation technique. Firstly, we compute the pre-estimation of this factor using equation (11), then using the pre-estimation result we directly obtain the amplitude of the clean speech component to refine the estimated Laplacian factor. Based on the MMSE theory, the amplitude of the clean speech component can be estimated using the following equation

$$\left|\hat{X}\right| = E\left\{\left|\hat{X}\right|\left|Y\right.\right\} = \frac{\int_{-\infty}^{\infty}\left|\hat{X}\right| p\{Y|X\}\, p\{X\}\, dX}{\int_{-\infty}^{\infty} p\{Y|X\}\, p\{X\}\, dX} \tag{16}$$

From equation (6) and (7), we get

$$p(X,Y) = \frac{1}{2a\sqrt{2\pi\sigma^2}}\exp(-\frac{|X|}{a} - \frac{(Y-X)^2}{2\sigma^2}) \tag{17}$$

and

$$p(X\,|\,Y) = \frac{p(X,Y)}{\int\limits_{-\infty}^{\infty} p(X,Y)dX} \tag{18}$$

Substituting (17), (18) into (16), we can easily obtain a novel estimation by

$$\begin{aligned}\left|\hat{X}\right| &= \frac{\int\limits_{-\infty}^{\infty}\left|\hat{X}\right|\exp(\frac{-|X|}{a} - \frac{|Y-X|^2}{2\sigma^2})dX}{\int\limits_{-\infty}^{\infty}\exp(\frac{-|X|}{a} - \frac{|Y-X|^2}{2\sigma^2})dX} \\ &= \frac{a\sqrt{\frac{8}{\pi}\psi}\exp(-\frac{\psi}{2} - \frac{\xi^2}{2\psi})}{\exp(\xi)\operatorname{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right) + \exp(-\xi)\operatorname{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right)} - \\ & a\frac{(\psi+\xi)\exp(\xi)\operatorname{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right) + (\psi-\xi)\exp(-\xi)\operatorname{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right)}{\exp(\xi)\operatorname{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right) + \exp(-\xi)\operatorname{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right)}\end{aligned} \tag{19}$$

where ξ, ψ and $\operatorname{erfc}(\cdot)$ are the same as equation (9). Then our algorithm for the Laplacian factor estimation can be obtained, which is included by equation (11) and (19).

As the estimated result can follow the shape of the real value without any simple delay, the estimated Laplacian factors using our approach can keep more accurate than the method in [9], and the performance of a speech enhancement system can also be improved accordingly.

4. EXPERIMENTAL RESULTS

In this section, the performance of the proposed approach is tested for speech enhancement, and compared to that of the original algorithm in [9]. The speech material used for tests consists of six sentences spoken by three males and three females. The number of samples per frame is *K*=128 with an overlap of 64 samples. The noise signals used in our evaluation are taken from http://spib.ece.rice.edu/, which include white noise (White), Pink noise (Pink), Buccaneer noise (Buccaneer), and Babble noise (Babble). The speech signal is sampled at 8 kHz and degraded by these noises at the SNR of 5dB, 10dB, and 15dB.

Firstly, the results of the two algorithms for speech enhancement are compared both in time and frequency domains by means of the waveform and spectrogram. Figure 1 to Figure 4 shows the waveform results of clean speech, noisy speech, and the enhanced speech corrupted by the Pink noise with 5 dB, and Figure 5 to Figure 8 gives the spectrograms of clean speech, noisy speech, and enhanced speech corrupted by the White noise with 10 dB. From the obtained results, it is apparent that our proposed approach has a better noise reduction capability, while keeping more of the speech signals energy unchanged than the original approach in [9].

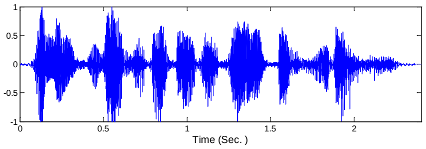

Figure1: Clean Speech

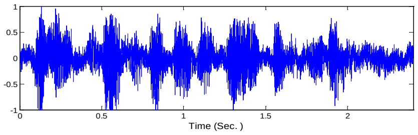

Figure2: Noisy Speech

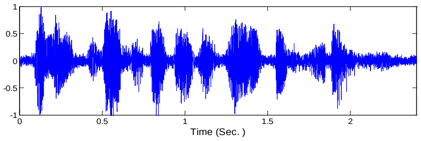

Figure3: The Enhanced Speech Using Method In [9]

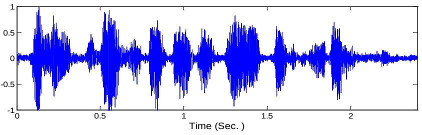

Figure4: The Enhanced Speech Using The Proposed Method

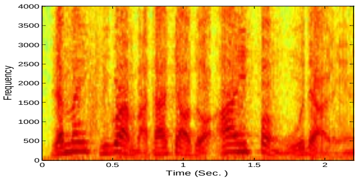

Figure 5: Clean Speech

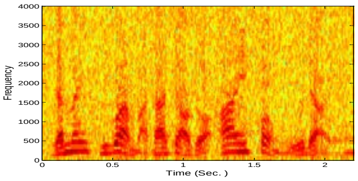

Figure6: Noisy Speech

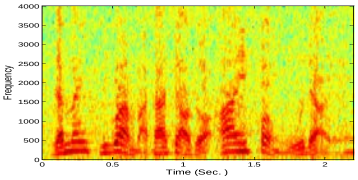

Figure7: The Enhanced Speech Using Method In [9]

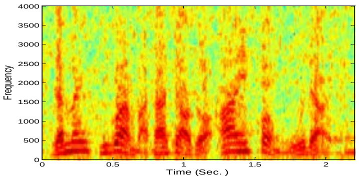

Figure8: Enhanced Speech Using The Proposed Method

The segmental SNR (SEGSNR) measure is adopted for the objective evaluation. For the segmental SNR, only frames with segmental SNR values greater than -10 dB and less than 35 dB are considered. Table 1 gives the output SEGSNR results of the enhanced speech signals obtained using the original and the proposed algorithm in various noise conditions and levels. From the table, we can observe that the proposed algorithm always has a higher SEGSNR as compared to the original algorithm under all tested environmental conditions.

Table 1: Comparison Of SEGSNR Of Enhanced Signals In Various Noise Conditions

Noise type	Input SNR	Output SEGSNR (dB)	
		Method in [9]	The proposed
White	5dB	6.7171	7.0737
	10dB	9.3519	9.6261
	15dB	11.8203	12.1225
Pink	5dB	6.2186	6.5337
	10dB	8.4528	8.7491
	15dB	11.6303	11.8832
Buccaneer	5dB	6.4059	6.6447
	10dB	8.7871	9.0994
	15dB	11.5438	11.7653
Babble	5dB	5.9859	6.2991
	10dB	8.6048	8.8542
	15 dB	11.8500	12.0710

5. CONCLUSION

Considering the speech enhancement problem using Laplacian priori in DCT domain, in this paper we present a novel approach for Laplacian factor estimation called two-step estimation technique. Based on the MMSE theory, the proposed estimator for Laplacian factor is refined by a second step to remove the bias of the DD approach, thus removing the reverberation effect. The experiment results have shown the improved performance of our method for speech enhancement in various noise conditions.

ACKNOWLEDGEMENTS

This work was supported by NSFC under Grant Nos. 61005021, 61201457 and A Project of Shandong Province Higher Educational Science and Technology Program under contract J12LN27.

REFERENCES:

[1] T. Inoue, H. Saruwatari, Y. Takahashi, K, Shikano, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 6, 2011, pp. 1770-1779.

[2] M. K. Hasan, M. S. A. Zilany, M. R. Khan, "DCT speech enhancement with hard and soft thresholding criteria", *Electronics Letters*, Vol. 38, No. 13, 2002, pp. 669-670.

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

[4] H, Ding, I. Soon, S. Koh, C. Yeo, "A spectral filtering method based on hybrid wiener filters for speech enhancement", *Speech Communication*, Vol. 51, No. 3, 2009, pp. 259-267.

[5] Y. A. Huang, J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 4, 2012, pp. 1256-1269.

[6] I. Y. Soon, S. N. Koh, C. K. Yeo, "Noisy speech enhancement using discrete cosine transform", *Speech Communication*, Vol. 24, No. 3, 1998, pp. 249-257.

[7] S. Gazor, W. Zhang, "Speech probability distribution", *IEEE Signal Processing Letters*, Vol. 10, No. 7, 2003, pp. 204-207.

[8] S. Gazor. "Employing Laplacian-Gaussian densities for speech enhancement", Proceedings of International Conference on Acoustic Speech Signal Processing, IEEE Conference Publishing Services, May 17-21, 2004, pp. 297-230.

[9] S. Ou, X. Zhao, Y. Gao, "Improved Laplacian factor estimation for speech enhancement", Proceedings of the 3rd International Conference on Wireless Communications, Networking and Mobile Computing, IEEE Conference Publishing Services, September 21-25, 2007, pp. 2911-2914.