14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

PARAMETRIC APPROACH FOR SPEECH DENOISING USING MULTITAPERS

Werayuth Charoenruengkit[1], Nurgun Erdol[2], and Tuncay Gunes[2]

International Business Machines (IBM)[1] and Florida Atlantic University (FAU)[2]
wcharoe@yahoo.com , erdol@fau.edu , tuncaygunes80@yahoo.com

ABSTRACT

Spectral estimation is a major component of obtaining high quality speech in many speech denoising techniques. Autoregressive spectral estimation using Multitaper Autoregressive data (ARMT) is a parametric approach that generates AR filter coefficients from multitaper autocorrelation estimates. The ARMT proves to be a best fit smooth curve to the mutitaper spectral estimates (MTSE) hence has very low high frequency bias and has even less variance than the standard MTSE. As such, ARMT is a smoother and less computationally intensive alternative to wavelet domain reduction (denoising) of the MTSE error. In this paper, the ARMT is used to derive the optimal gain parameters in the signal subspace approach to reducing environmental noise. Objective measures and informal listening tests demonstrate that results are indistinguishable from its successful predecessor that uses the non-parametric approach for speech denoising.

1. INTRODUCTION

Speech denoising methods of spectral subtraction, minimum mean-square error (MMSE) [2] and signal subspace [1][2][5][8][13] are all non-parametric. They use direct signal spectrum estimates related to the periodogram which trade off spectral leakage and bias with very high variance. Multitaper Spectral Estimation (MTSE) exchanges variance and bias and achieves statistical stability by averaging direct spectral estimates over multiple tapers [12].

Recent work [8] has shown that speech enhancement by reduction of environmental noise can be improved by using MTSE that has been refined and smoothed by wavelet thresholding techniques [14]. The underlying idea behind these techniques is to represent the log periodogram as "signal" plus the "noise," where the signal is the true spectrum and "noise" is the estimation error. Reducing spectral estimation errors leads to improvements in SNR estimates in noisy speech frames [8] which are used in signal subspace methods [1] for speech enhancement. Since the MTSE has no parameters characterizing the nature of speech, adjustment to its performance, for example by incorporating perception factors, becomes difficult. In [4], it has been shown that MTSE can be parameterized successfully, resulting in an autoregressive (AR) spectral estimate that is smoother than is the result of wavelet shrinkage and has all the statistical advantages of the MTSE. The ARMT method is also a low-computation alternative to speech enhancement proposed in [8].

In this paper, we show that AR coefficients derived from the non-parametric MT autocorrelation (MTAC) estimates [4] generate a smooth tight-fitting curve (ARMT) to the MTSE. As such, it has very low high frequency bias and has even less variance than the standard MTSE. This proves to be [4] a method of reducing spectral estimation errors that is computationally less intensive than the wavelet thresholding method [14] employed for the same purpose. We then use the ARMT to derive the gain parameters of the optimal estimator [8] of enhanced speech.

In the next section, we review the subspace method [1] of signal enhancement. In section 3, we develop the algorithm of adjusting gains derived from the ARMT parameters. In section 4, we show results of speech denoised from additive white Gaussian noise, additive colored (AR) noises, and real automotive noise. Conclusions follow in Section 5.

2. SHORT TIME SPECTRAL AMPLITUDE ESTIMATOR

If X and D are the $N \times 1$ vectors containing spectral components of the clean speech signal vector x and the noise vector d , then the Fourier transform of the noisy speech signal vector y can be written as $F^H y = Y = X + D = F^H x + F^H d$, where F^H is a N-point discrete Fourier Transform matrix. Letting $\hat{X} = GY$ be the linear estimator of X , the error is defined as $\varepsilon = \hat{X} - X = \varepsilon_X + \varepsilon_D$ where speech distortion $\varepsilon_X = (G - I)X$ and residual noise $\varepsilon_D = G \cdot D$. The estimation operator $\mathbf{G}$ is chosen to minimize both $\bar{\varepsilon}_X^2 = E(\varepsilon_X^H \varepsilon_X)$ and $\bar{\varepsilon}_D^2 = E(\varepsilon_D^H \varepsilon_D)$. The minimization problem can be solved with the following constrained optimization problem:

$$\min_{G} \bar{\varepsilon}_X^2 \quad \text{Subject to} \ : \ \frac{1}{N}\bar{\varepsilon}_D^2 \le c \tag{1}$$

where c is an integer.

The optimal G obtained from using Kuhn-Tucker technique [14] satisfies the following equation

$$G(F^H R_X F + \mu F^H R_D F) = F^H R_X F \tag{2}$$

where μ is the Lagrange multiplier.

If R_X and R_D are Toeplitz, $F^H R_X F$ and $F^H R_D F$ are asymptotically diagonal [6] with diagonal element equal to power spectrum components $S_x(w)$ and $S_d(w)$ of the clean speech vector x and noise vector d, respectively. As a result, matrix G is also diagonal with k^{th} diagonal $g(k)$ and given by

$$g(k) = \frac{S_x(k)}{S_x(k) + \mu \cdot S_d(k)} = \frac{\gamma_{prio}(k)}{\gamma_{prio}(k) + \mu} \tag{3}$$

where $\gamma_{prio}(k) = \dfrac{S_x(k)}{S_d(k)}$ is the a priori SNR at frequency k and μ is a control parameter dependent to the noise level.

Hu [8] estimates $\gamma_{prio}(k)$ using the non-parametric of Multitaper Wavelet Denoising (MTW) spectral estimation applied to estimate $S_x(w)$ and $S_d(w)$, and controls the suppression level with the Lagrange multiplier based on posterior SNR. In the next section, we provide an alternative method of calculating this gain function using parametric method derived from MTSE. The proposed parametric approach reduces the computational complexity while also achieving low variance estimates. Applying gain function g(k) to the noisy speech spectrum results an optimal enhanced speech

3. POWER SPECTRAL ESTIMATION

3.1. Non-parametric approach for spectral estimation

Spectral estimation based on DFT is the most popular approach due to its simplicity. However, it introduces a large bias and variance as well as spectral leakage. Carefully chosen tapers such as Hamming windows can help reduce leakage and bias. However, they do not reduce the variance of the spectral estimate [9]. Multitaper spectral estimate reduces variance in exchange for a slight increase in bias. The bias reduction estimator by a factor of 1/K [12] can be obtained from the average of the direct spectral estimates of data multiplied with K tapering windows. The tapers may be chosen to be pairwise orthogonal and are properly designed to prevent leakage. Most commonly used tapers are discrete prolate spheroidal sequences (dpss) [12] or sinusoidal window sequences [11], and other low-leakage combinations are given in [3]. The multitaper spectrum estimator with K tapers is given by

$$S_{mt}(w) = \frac{1}{K}\sum_{k=0}^{K-1} \left| X_k(w) \right|^2 \tag{4}$$

where

$$X_k(w) = \frac{1}{N}\sum_{n=0}^{N-1} x[n] v_k[n] \exp(-jwn) \tag{5}$$

$X_k(w)$ is the discrete-time Fourier transform (DTFT) of data $x[n]$ of a random process x_n multiplied by window sequence $v_k[n]$.

3.2. Proposed parametric approach

Linear predictive coding (LPC) is a well-known and effective technique for estimating the time-varying resonances of the vocal tract from speech signals. With this system, it is possible to independently control the strengths and frequencies of each formant. For an AR process, a signal $x[n]$ is assumed to be predictable by its past value and input $w[n]$ with a gain G and is given by

$$x[n] = \sum_{k=1}^{p} a_k x[n-k] + Gw[n] \tag{6}$$

This relationship can be expressed in terms of a transfer function as

$$H_{AR}(z) = \frac{X(z)}{W(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{7}$$

In [3][4], it was suggested that the non-parametric MTSE given in (4) and (5) may be expressed as

$$\hat{S}(\omega) = \iint_{\gamma,\xi} d\gamma d\xi X(\gamma) X^*(\xi) A_{K,N}(\omega - \gamma, \omega - \xi) \tag{8}$$

where

$$A_{K,N}(\gamma, \xi) = \frac{1}{K}\sum_{k=0}^{K-1} V_k(\gamma) V^*{}_k(\xi) \tag{9}$$

$V_k(w)$ are the DTFTs of the tapers. In time (or index) domain, the above relationship can be expressed by

$$\hat{r}[n] = \sum_{m=0}^{N-1} x[m] x[m-n] \alpha_{K,N}[m,n] \tag{10}$$

where

$$\alpha_{K,N}[m,n] = \frac{1}{K}\sum_{k=0}^{K-1} v_k[m] v_k[m-n] \tag{11}$$

$\hat{r}[n]$ is the IDFT of $\hat{S}(\omega)$. Given (10) and (11), the AR filter coefficients in (7) can be estimated as follows:

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r} \tag{12}$$

where $\mathbf{a} = [a_1\, a_2 ... a_p]^T$, R is the matrix of autocorrelation estimate $\hat{r}[n,m]$, and $\mathbf{r} = [\hat{r}[1,0]\hat{r}[2,0]...\hat{r}[p,0]]^T$. The solution can be easily computed with the Levinson-Durbin algorithm [7] . The AR coefficients obtained from (12) are used to generate AR spectra in (7) , which is used to estimate the optimal gain estimation in (3).

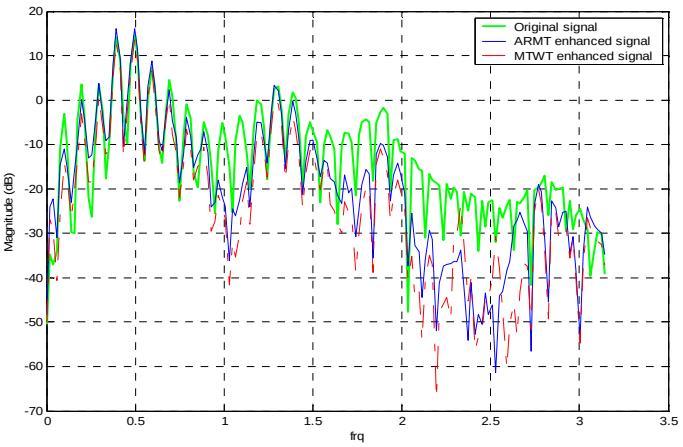

Fig. 1. Periodogram of the enhanced speech obtained from MTWT and MTAR compared with the original clean speech.

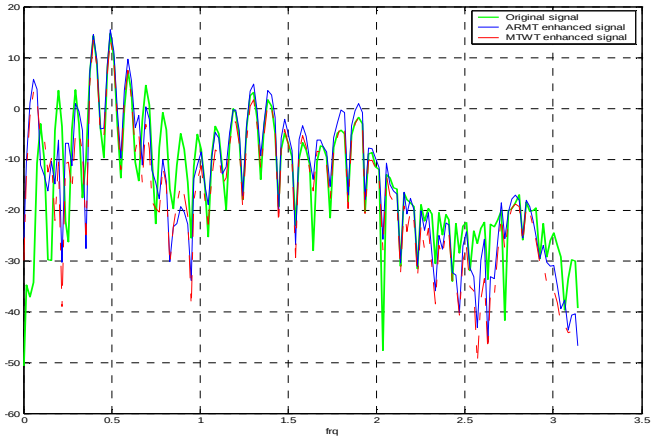

Fig. 2. Periodogram of the enhanced speech obtained from MTWT and MTAR, with pre-emphasis filter, compared with the clean speech.

4. RESULTS

The experiment is conducted by using a clean speech additively mixed with 4 noise types: white noise, AR-3 process noises, AR-4 process noise, and a recorded car noise. The AR process noise is created by filtering white noise with the AR filter. The AR-4 filter is derived from an arbitrary chosen speech frame where poles are in the location obtained from speech signal. The AR-3 filter is selected from a stable AR-filter where the poles are not in the region of normal speech signal [10]. The audio is sampled at 8K Sample-Hz. The estimated spectral gains and denoised speech obtained from the proposed technique are compared to those obtained from the MTW and the auditory model incorporating spectral subtraction (AD-SS) [13].

TABLE I
COMPARATIVE PERFORMANCE FOR SPEECH ENHANCEMENT IN TERMS OF GLOBAL SNR AND SEGMENTAL SNR (SNR/SNR_SEG) FOR A SENTENCE ADDITIVELY MIXED WITH DIFFERENT NOISE TYPES

	White noise	AR-3 Noise	AR-4 Noise	Car Noise
Noisy Speech	-6.74/-8.17	-6.76/-8.17	-6.6/-8.10	-6.86/-7.94
AD-SS	3.74/-3.01	3.80/-3.55	3.46/-3.63	4.90/-2.69
MTW	4.51/-3.30	4.49/-3.34	4.40/-3.30	7.50/-2.21
ARMT	4.51/-3.34	4.49/-3.50	4.38/-3.37	7.42/-2.19

TABLE II
COMPARATIVE PERFORMANCE FOR SPEECH ENHANCEMENT IN TERMS OF MODIFIED SEGMENTAL SNR FOR 10 SENTENCES (AVERAGE SEGMENTAL SNR/AVERAGE LOWER BAND SNR, AVERAGE UPPERBAND SNR)

	Male Speaker	Female Speaker
Noisy Speech	-2.20 / 0.95 / -4.5	-1.58 / -0.19 / -7.26
MTW	1.82 / 2.89 / -1.69	0.96 / 1.73 / -2.97
ARMT	1.85 / 2.90 / -1.78	0.98 / 1.75 / -3.0
AD-SS	1.67 / 2.48 / -1.59	0.65 / 1.25 / -2.56

The initial result indicates that the proposed technique and the MTW may not estimate the speech spectrum well in the high frequency band as the result shown in Fig 1. Additional improvement can be obtained by applying a pre-emphasis filter to the noisy signal before the AR coefficient estimation. The improvement result is shown in Fig 2 where the enhanced speech tracks closer to the clean speech signal at the high frequency band after applying a pre-emphasis filter. The improvement in the high frequency band is not subjected to the noise types used in the experiment.

The objective test is calculated by using the Global SNR and a modified average Segmental SNR (SNR_{seg}^{ave}). The SNR_{seg}^{ave} are implemented in the frequency band so that the SNR can be calculated within the interested frequency. It can be obtained by averaging Segmental SNR in the frequency domain over M speech frames as follows:

$$SNR_{seg}^{ave} = 10 * \log 10(10^{(\frac{1}{M}\sum_{k=0}^{M-1} SNR_{seg}^{k})-1})$$

$$where\ SNR_{seg}^{k} = \log 10(1+\frac{\sum_{w=0}^{N-1}\left|X^{k}(w)\right|^{2}}{\sum_{w=0}^{N-1}(\left|X^{k}(w)\right|-\left|\hat{X}^{k}(w)\right|)^{2}}) \quad (13)$$

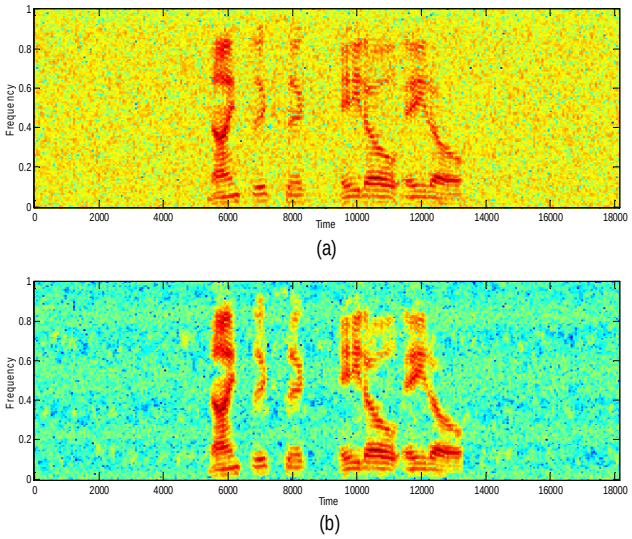

Fig. 3. Speech Spectrograms (a) Noisy speech signal corrupted by Gaussian white, (b) Denoised speech signal using the ARMT method.

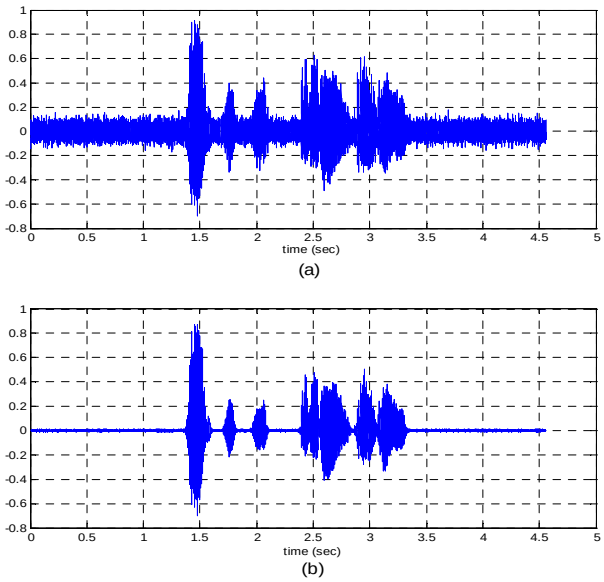

Fig. 4. Speech Signal (a) Noisy speech signal orrupted by Gaussian white, (b) Denoised speech signal using the ARMT method.

Table I presents the Global SNR and SNR_{seg} of the speech signal corrupted by different noise types and of the denoised speech signal obtained from the different denoising techniques. The highest SNR improvement is obtained from denoising car-type noise. This may contribute to the properties of the car noises that have most energy at the lower frequency band. These results indicate that the proposed technique may work well in the lower frequency bands.

More investigation on the performance in lower band SNR and upper band SNR are conducted on the noisy speech signals obtained from additively mixing with white Gaussian noise. The SNR results are average from 5 male speakers and 5 female speakers. Table II shows the performance of the denoising techniques in the lower-band frequency (0-2 kHz) and in the higher band (2 kHz-4 kHz) for both male and female speakers. In the lower-band frequency, the SNR obtained from both ARMT and MTW are higher than those obtained from AD-SS. However, the proposed method little outperforms MTW. The great improvement in quality is mainly obtained with the male speakers. This may contribute to the fact that the proposed spectral estimation of speech signal performs best with low pitch speech signal spoken by male speakers.

An informal listening test is conducted by asking 10 listeners to give a score for each sentence they listen, in terms of being more natural and having less distortion. The results from the subjective test indicate that the enhanced speech obtained from MTW and ARMT are almost indistinguishable and are better than those obtained from the AD-SS. In accordance with Erdol and Gunes in [4], the spectral estimate obtained from ARMT is smoother to that from MTW but nearly identical. As a result, the enhanced speech is almost indistinguishable between these two approaches.

Since human perception to the audio distortion is more sensitive in the lower-band frequency, the AD-SS is less preferable than those obtained from the proposed method, as also supported by the objective test results shown in table II. Fig 3 presents the spectrogram of noisy speech obtained from additive white Gaussian noise at 1 dB SNR and the spectrogram of denoised speech using ARMT. Fig 4 presents the same signals in the time domain. The distortion and residual noise in the proposed technique is much less than those obtained from the AD-SS. Moreover, no disturbing musical noise is observed in the denoised speech obtained from the proposed technique.

5. CONCLUSION

Quantitative and qualitative tests show that speech enhancement using proposed ARMT spectral estimates has better quality than the AD-SS, and is as good as the MTW. A

greater improvement is obtained in the lower frequency region where most speech information is located. In addition, the computation time of ARMT is much less than MTW. One of the computational simplifications implemented in the ARMT is that the autocorrelation computation of windows in equation (9) can be pre-computed off-line whereas the multiple tapers used in MTW computation has to be computed during the run-time. In addition, it is not necessary to perform Wavelet Denoising as in Hu [8] since the spectrum obtained from the ARMT incorporates spectrum smoothness as the nature of AR process. ARMT also yields parameters for later modification of the poles and zeros or spectral features to achieve the optimal result and higher speech quality. Further investigation is possible to reduce computation complexity by calculating spectral gains using other forms of parameters instead of the direct computation in the spectral domain.

6. ACKNOWLEDGEMENT

This material is partially supported by grant NNL05AA02G from the National Aeronautics and Space Administration (NASA).

7. REFERENCES

[1] Ephraim, Y. and Van Trees, H. L. , “A signal subspace approach for speech enhancement,” IEEE Trans. Speech and Audio Processing, vol. 3, pp. 251-266, July 1995.

[2] Ephraim Y. and Malah, D., “Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” IEEE Trans. on Acoust., Speech, Signal Processing, vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[3] Erdol N., “Multitaper spectral estimation: A generalized parametric approach”. EUSIPCO 2005, 13th European Signal Processing Conference, September 4-8, 2005, Antalya, Turkey

[4] Erdol N. and Gunes T., “Multitaper covariance estimation and spectral denoising”. Proceedings of the 39th Asilomar Conference on Signals, Systems and Computers, 2005

[5] Boll, S.F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, April 1979.

[6] Gray, R. M., "On the asymptotic eigenvalue distribution of Toeplitz matrices," IEEE Transactions on Information Theory, vol. 18, pp. 725-730, Nov. 1972.

[7] Haykin, S. Adaptive Filter Theory., Prentice Hall, 1996.

[8] Hu, Y. and Loizou, P., "Speech enhancement by wavelet thresholding the multitaper spectrum," IEEE Transactions on Speech and Audio Processing, 12(1), 59-67, 2004

[9] Kay, S. Modern Spectral Estimation: Theory, Prentice Hall, 1998.

[10] Oppenheim, A.V. and Schafer, R.W.. Discrete-time signal processing. Prentice-Hall, 1989.

[11] Riedel, K.S. and Sidorenko, A., “Minimum bias multiple taper spectral estimation,” IEEE Transactions on Signal Processing, 43, no. 1, 188-195, 1995.

[12] Thomson, D.J., “Spectrum estimation and harmonic analysis,” Proceedings of the IEEE, 70, 1055-1096, 1982.

[13] Virag, N., “Single channel speech enhancement based on masking properties of the human auditory system,” IEEE Transactions on Speech and Audio Processing, 7, 126–137, 1999.

[14] Walden, A.T., Percival, D.B. and E.J., McCoy, “Spectrum estimation by wavelet thresholding of multitaper estimators,” IEEE Trans. Signal Processing, 46, 3153-3165. 1998.