TEMPORAL MODELLING AND KALMAN FILTERING OF DFT TRAJECTORIES FOR ENHANCEMENT OF NOISY SPEECH

Esfandiar Zavarehei, Saeed Vaseghi, Qin Yan

School of Design and Engineering, Brunel University, London UK
{esfandiar.zavarehei, saeed.vaseghi, qin.yan}@brunel.ac.uk

ABSTRACT

This paper presents a time-frequency estimator for enhancement of noisy speech in the DFT domain. The time-varying trajectories of the DFT of speech and noise in each channel are modeled by low order autoregressive processes incorporated in the state equation of Kalman filters. The parameters of the Kalman filters are estimated recursively from the signal and noise in DFT channels. The issue of convergence of the Kalman filters to noise statistics during the noise-dominated periods is addressed and a method is incorporated for restarting of Kalman filters after long periods of noise-dominated activity in each DFT channel. The performance of the proposed method is compared with cases where the noise trajectories are not explicitly modeled. Evaluations show that the proposed method results in substantial improvement in perceived quality of speech.

1. INTRODUCTION

Speech enhancement improves the quality and intelligibility of voice communication for a range of applications including mobile phones, teleconference systems, hearing aids, voice coders and automatic speech recognition. Among different solutions proposed for enhancement of noisy speech, restoration of short-time speech spectrum has been extensively studied [1][2]. This approach is normally based on estimation of the short time spectral amplitude (STSA) of the clean speech using an estimate of the signal-to-noise ratio (SNR) at each frequency. The effect of phase distortion is assumed to be inaudible.

An alternative to estimation of the STSA is the estimation of the real and imaginary components of the DFT of the clean speech. The MMSE estimation of the DFT components with Gaussian priors, leads to the well-known Wiener filter solution [3] while the MMSE estimation of the STSA within the same set of Gaussian assumptions results in Ephraim's noise suppression method [1]. In recent years Martin has proposed the use of Gamma and Laplacian distributions for modeling the real and imaginary components of the DFT of speech [3].

Speech enhancement methods often assume that the spectral samples are independent identically distributed (IID) samples across frequency and time dimensions. However, there seems to be an apparent contradiction [4]; these same methods that start with the IID assumption, often also use the assumption of the dependency of successive frames for the calculation and smoothing of some key speech parameters such as the SNRs [1][3][5].

The modeling and utilization of the time-varying trajectory of speech and noise spectrum is the main focus of this paper. In this paper, the temporal trajectory model of the DFT of speech and noise are used in a more rigorous mathematical framework for a more reliable estimation of speech spectra. The use of Gaussian priors lends itself to application of Kalman filter for modeling the temporal trajectories of the DFT of speech. A set of AR models are incorporated in Kalman filters for adaptive estimation and modeling of the temporal trajectories of the DFT of the speech and noise signals.

The rest of this paper is organized as follows. Section 2 discusses the modeling of the samples of the temporal trajectories of DFT components. In Section 3 the Kalman estimator of DFT trajectories is introduced. In Section 4 the empirical issues and the parameter estimation of the new estimator are discussed. In section 5 the evaluation results are compared with other methods of speech enhancement. Conclusions are drawn in Section 6.

2. MODELLING DFT TRAJECTORIES

In this section the temporal dependency and predictability of the trajectory of the DFT components are examined. The level of correlation between successive temporal samples of DFT components varies for different frequencies as well as different phonemes (i.e. along time and frequency). Moreover, the probability distributions of DFT components are strongly dependent on the frequency channel and the phoneme under study. Figure 1 illustrates the distribution of DFT components of channel 26 (1000 Hz) for phoneme /*ah*/. The data is obtained from 130 sentences spoken by a male speaker selected randomly from the Wall Street Journal (WSJ) database. It is evident from Figure 1 that the peak of the histogram is modeled better with a Gamma distribution while the sides tend to fit a Gaussian distribution.

Table 1 shows the average symmetric Kullback-Leibler distance (SKLD) [6] between histograms and parametric distributions. These results show that, on average, Gamma distribution models the distribution of DFTs of speech better than Gaussian distribution. This is observed from the SKLD of speech with parametric distributions. It is also observed that most noise types

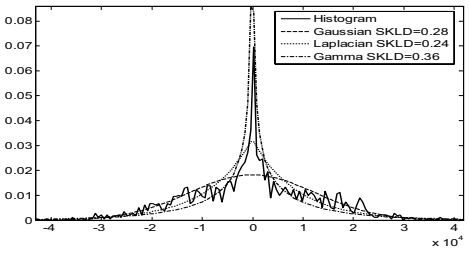

Figure 1: Normalized histogram of ST-DFT components for channel 26 (1 kHz), Phoneme /*ah*/

Table 1. Average SKLD between the histograms and different parametric distributions for speech (averaged over all phonemes/frequency channels for 130 labeled sentences spoken by a male speaker) and different noise types

Distribution	Gaussian	Laplacian	Gamma
Speech	0.81	0.62	0.56
Car noise	0.04	0.10	0.85
Train noise	0.15	0.05	0.22
Babble noise	0.69	0.51	0.46
Helicopter fly-by noise	0.12	0.15	0.59
White Gaussian	0.01	0.22	0.83

have a reasonably low SKLD with the Gaussian distribution. However, as often, a compromise, between the complexity and the mathematical tractability of the model, suggests the use of Gaussian distribution and Kalman filters for modeling the temporal trajectories of DFT. The real part of the DFT of clean speech, $S_r(n)$, can be modeled using an AR process:

$$S_r(n) = \sum_{k=1}^{N} a_k(n) S_r(n-k) + e_r(n) \tag{1}$$

where $S_r(n)$ is the real part of the DFT of clean speech at frame n of an arbitrary frequency channel, $a_k(n)$ is the k^{th} AR coefficient at the n^{th} frame of the same frequency channel, $e_r(n)$ is the corresponding estimation error and N is the model order. Moreover, it is assumed that $S_r(n)$ is a stationary process within the prediction period. Assuming Gaussian distributions for DFT components, the MMSE linear predictor (LP) coefficients of Equation (1) can be obtained using Yule-Walker equation:

$$\mathbf{a}(n) = \left(\mathbf{R}_{s_r}(n)\right)^{-1} \mathbf{r}_{s_r}(n) \tag{2}$$

where $\mathbf{R}_{s_r}(n)$ and $\mathbf{r}_{s_r}(n)$ are the autocorrelation matrix and vector of the real part of speech DFT, $\mathbf{S}_r(n)=[S_r(n), \ldots S_r(n\text{-}L\text{+}1)]^T$, respectively and $\mathbf{a}(n)$ is the AR coefficient vector at frame n. A similar equation stands for the imaginary component of the DFT. The speech frame length, overlap size and the LP order should be carefully chosen to comply with the stationarity assumption of Equation (1), that is between say 20-40 ms.

Figure 2 illustrates the correlation coefficients between delayed samples of the DFT of noise and speech signals, averaged over all frequency channels. Note however, that while the correlation coefficient may be negative, it is the absolute value which shows the level of correlation. It is evident that although, due to the frame overlap, there is a correlation between successive samples of DFT of noise, this does not vary much with the noise type and is less than that of speech. The shift-size used in Figure 2 is 5ms and the frame size is 25ms which experimentally proved to result in good noise reduction.

3. KALMAN DFT TRAJECTORY RESTORATION

This section presents the formulation of Kalman filters for restoration of DFT trajectories. It is assumed that the clean speech signal $s(t)$ is contaminated by the additive background noise $d(t)$ uncorrelated with the speech signal. The noisy speech signal $x(t)$ is modelled as:

$$x(t) = s(t) + d(t) \tag{3}$$

where t denotes time. For each frequency channel Equation (3) is rewritten in DFT domain as:

$$X_r(n) + jX_i(n) = \left(S_r(n) + D_r(n)\right) + j\left(S_i(n) + D_i(n)\right) \tag{4}$$

where the subscripts r and i represent the real and imaginary parts of DFT respectively and n denotes frame index. It is assumed that

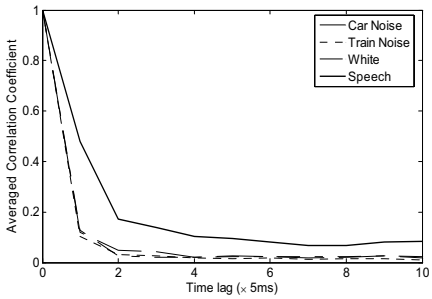

Figure 2: Averaged absolute correlation in ST-DFT trajectories.

the real and imaginary parts of the DFT are independent and have Gaussian distributions. The independency assumption of the real and imaginary components is verified from a study of the scatter plots of the real and imaginary parts of the DFT coefficients of clean speech [3][7]. The real part of the DFT of noise, $D_r(n)$, is modeled using an AR model as:

$$D_r(n) = \sum_{k=1}^{M} b_k(n) D_r(n-k) + g_r(n) \tag{5}$$

where $D_r(n)$ is the real part of the DFT of noise at frame n of an arbitrary frequency channel, $b_k(n)$ is the k^{th} AR coefficient at the n^{th} frame of the same frequency channel, $g_r(n)$ is the corresponding estimation error which has a variance of $\sigma_{g_r}^2(n)$ and M is the model order. Following straight-forward algebra manipulation, equations (1), (4) and (5) for the real part may be represented in canonical form:

$$\mathbf{X}_r(n) = \mathbf{A}_r(n)\mathbf{X}_r(n-1) + \mathbf{G}_c\mathbf{E}_r(n) \tag{6}$$

$$X_r(n) = \mathbf{H}_c\mathbf{X}_r(n) \tag{7}$$

where the state vector $\mathbf{X}_r(n)$ is defined as:

$$\mathbf{X}_r(n) = \left[\mathbf{S}_r^T(n) \quad \mathbf{D}_r^T(n)\right]^T \tag{8}$$

$$\mathbf{S}_r(n) = \left[S_r(n-N+1) \quad \cdots \quad S_r(n)\right]^T \tag{9}$$

$$\mathbf{D}_r(n) = \left[D_r(n-M+1) \quad \cdots \quad D_r(n)\right]^T \tag{10}$$

where $\mathbf{S}_r$ and $\mathbf{D}_r$ are speech and noise state vectors respectively. The transition matrix $\mathbf{A}_r(n)$ is given by:

$$\mathbf{A}_r(n) = \left[\begin{array}{c|c} \mathbf{F}_r(n) & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{B}_r(n) \end{array}\right] \tag{11}$$

$\mathbf{F}_r(n)$ and $\mathbf{B}_r(n)$ are speech and noise transition matrices respectively:

$$\mathbf{F}_r(n) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_N(n) & a_{N-1}(n) & a_{N-2}(n) & \cdots & a_1(n) \end{bmatrix} \tag{12}$$

$$\mathbf{B}_r(n) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ b_M(n) & b_{M-1}(n) & b_{M-2}(n) & \cdots & b_1(n) \end{bmatrix} \tag{13}$$

$\mathbf{E}_r(n)$ is the AR error vector of noise and speech and $\mathbf{H}_c$ and $\mathbf{G}_c$ are constant vectors defined below:

$$\mathbf{E}_r(n) = \left[e_r(n) \quad g_r(n)\right]^T \tag{14}$$

$$\mathbf{G}_c = \begin{bmatrix} \mathbf{U}(N) & \mathbf{0} \\ \mathbf{0} & \mathbf{U}(M) \end{bmatrix} \quad (15)$$

$$\mathbf{H}_c = \begin{bmatrix} \mathbf{U}^T(N) & \mathbf{U}^T(M) \end{bmatrix} \quad (16)$$

where $\mathbf{U}(N)$ is a $N\times1$ vector defined as:

$$\mathbf{U}(N) \triangleq \begin{bmatrix} \overbrace{0 \quad \cdots \quad 0}^{N-1} & 1 \end{bmatrix}^T \quad (17)$$

A *prediction* of the state vector is obtained from the previous state vector using the transition matrix $\mathbf{A}(n)$ as:

$$\hat{\mathbf{X}}_r^-(n) = \mathrm{E}\left\{\mathbf{X}_r(n) \,|\, \hat{\mathbf{X}}_r(n-1)\right\} = \mathbf{A}_r(n)\hat{\mathbf{X}}_r(n-1) \quad (18)$$

where $\hat{\mathbf{X}}_r(n-1)$ is the *estimate* of $\mathbf{X}_r(n\text{-}1)$. As $e_r(n)$ and $g_r(n)$ are orthogonal to $\hat{\mathbf{X}}_r(n-1)$ and each other, the *prediction* error covariance matrix is calculated as:

$$\mathbf{P}_{rc}^-(n) = \mathbf{A}_r(n)\mathbf{P}_{rc}(n-1)\mathbf{A}_r^T(n) + \mathbf{G}_c\mathbf{\Lambda}(n)\mathbf{G}_c^T \quad (19)$$

$\mathbf{\Lambda}(n)$ is a 2×2 matrix defined as:

$$\mathbf{\Lambda}(n) \triangleq \begin{bmatrix} \sigma_{e_r}^2(n) & 0 \\ 0 & \sigma_{g_r}^2(n) \end{bmatrix} \quad (20)$$

and $\mathbf{P}_{rc}(n-1)$ is the state *estimation* error covariance matrix. Note that the innovation here is the difference between the predicted noisy signal and the observed noisy signal as according to Equation (7) there is no "noise" added to $\mathbf{H}_c\mathbf{X}(n)$. Incorporating the innovation in the current noisy observation, the optimum estimate of the state vector is calculated as:

$$\hat{\mathbf{X}}_r(n) = \hat{\mathbf{X}}_r^-(n) + \mathbf{K}_{rc}(n)\left(X_r(n) - \mathbf{H}_c\hat{\mathbf{X}}_r^-(n)\right) \quad (21)$$

where $\mathbf{K}_{rc}(n)$ is the Kalman gain vector:

$$\mathbf{K}_{rc}(n) = \mathbf{P}_{rc}^-(n)\mathbf{H}_c^T\left[\mathbf{H}_c\mathbf{P}_{rc}^-(n)\mathbf{H}_c^T\right]^{-1} \quad (22)$$

Note that $\mathbf{H}_c\mathbf{P}_{rc}^-(n)\mathbf{H}_c^T$ is a scalar value. The *estimation* error covariance of this estimate, $\mathbf{P}_{rc}(n)$, is obtained as:

$$\mathbf{P}_{rc}(n) = \left[\mathbf{I} - \mathbf{K}_{rc}(n)\mathbf{H}_c\right]\mathbf{P}_{rc}^-(n) \quad (23)$$

The same set of equations holds for the imaginary component of all frequency channels with nonzero imaginary parts. The estimated clean speech DFT is the by-product of the $\mathbf{X}$ in Equation (21).

4. PARAMETER ESTIMATION

As the autocorrelation of the DFT trajectories of clean speech is not available for estimation of AR parameters in Equation(2), the autocorrelation vector obtained from the past restored samples is used. That is:

$$\hat{\mathbf{a}}(n) = \left(\hat{\mathbf{R}}_{sr}(n-1)\right)^{-1} \times \hat{\mathbf{r}}_{sr}(n-1) \quad (24)$$

The autocorrelation vector and matrix are calculated from the past L=8 samples (with a shift-size of 5ms this is equivalent to 40ms).

An implementation issue arises from the feedback of restored speech for calculation of AR parameters using Equation (24). During long (typically >200ms) noise-only periods, where the variance of the noisy signal is equal to that of noise, the recursive solution given by Equations (19) and (22), results in convergence of the output of Equations (21) towards zero which consequently decreases the variance of prediction error, $\sigma_{e_r}^2(n)$, towards zero. In other words, the Kalman filters speech output converges to zero during noise-only periods. At the beginning of the speech signal, just after a long noise-only period, due to the suppression of noise and the absence of speech the prediction of the DFT trajectories will be very small with a consequently small prediction error variance, $\sigma_{e_r}^2(n)$, which results in a high weight for the prediction of the state vector (very small Kalman gain) and zeroing of the output speech signal. In order to prevent the consequent zeroing of speech following a long period of speech inactivity the value of $\sigma_{e_r}^2(n)$ needs to be revived from zero at the beginning of speech active periods. This is achieved by ensuring that values of $\sigma_{e_r}^2(n)$ will not be less than a dynamic threshold which is a fraction of the noisy signal energy at each time-frequency bin. That is:

$$\hat{\sigma}_{e_r}^2(n) = \max\left(\sigma_{e_r}^2(n), \alpha^2|X(n)|^2\right) \quad (25)$$

This limits the prediction error variance to a small portion of the instantaneous power spectrum of noisy speech. Equation (25) implies that the DFT trajectories can be only predicted with a limited precision, i.e. the prediction error variance cannot be smaller than a threshold proportional to the variance of the noisy speech. Very small values for α proved to be sufficient for reviving the converged trajectories of $\sigma_{e_r}^2(n)$ and the signal at the beginning of speech activity (e.g. α=0.07).

In order to obtain the AR models of the DFT trajectories of noise for each frequency channel, the autocorrelation of the DFT trajectories are obtained and smoothed during the noise-only periods. These autocorrelation vectors are obtained using L samples of the real and imaginary components separately and then averaged for each time step. That is, the same AR model is used for the real and imaginary components of each channel of noise.

5. EVALUATION RESULTS

The evaluation of the performance of DFT-Kalman filter with correlated noise model (DFTKCN) described in section 3, for enhancement of speech signals corrupted by background noise is carried out using subjective and objective measures. Various types and levels of noise are added to the speech signals selected from the WSJ speech database. The noisy signals are segmented using 25ms hamming windows with a shift size of 5ms. The car noise signal is recorded by our colleagues in a 3-series BMW at 70 Mph in a rainy day and the train noise is recorded in a moving train. The parameters used in Kalman method are: Autocorrelation length L=8, LP orders N=4 and M=2 and α=0.07.

5.1. Mean Opinion Score (MOS)

A set of twenty sample sentences are drawn from WSJ database and contaminated by car noise and train noise at two different SNRs, 0dB and 10dB. The resulting noisy speech sentences are then de-noised using four different methods: (i) parametric spectral subtraction (PSS) [2], (ii) MMSE log-STSA [5], (iii) DFT-Kalman filter with uncorrelated noise model [8] (DFTKUN) and (iv) DFTKCN. Note that in the first two methods decision-directed method is used for tracking the *a priori* SNR [1]. Ten trained listeners were asked to score the quality of the resulting output signals from 1 to 5, based on the perceptual ease of understanding (intelligibility) and the comfort of listening (less annoying noise). The mean opinion score results are presented in Table 2. The results of Table 2 show that the Kalman filter outputs are preferred by the listeners. As often, the extent of validity of these results is limited by the number of listeners and test sentences used.

Table 2: Mean opinion score results

SNR	Noise	DFTKUN	DFTKCN	MMSE	PSS	Wiener
0dB	Car	3.7	3.8	3.5	3.4	3.2
	Train	2.7	2.9	2.0	2.0	2.1
10dB	Car	4.5	4.7	4.6	4.4	4.2
	Train	3.7	3.9	3.7	3.3	3.5

Table 3: The correlation coefficient ρ of MOS and objective evaluation results

	PESQ	LLR	ISD	Kullback	SegSNR	SNR
ρ	0.86	-0.69	-0.61	-0.45	0.24	0.07

5.2. Objective Evaluation

From a number of different speech quality and distortion measures applied to the restored sample speech sentences of section 5.1, six are listed in Table 3. The correlation coefficient of each distortion measure with MOS was calculated and the three most correlated distortion measures were chosen for further objective evaluation of the performance of different methods. Table 3 summarizes the correlation coefficients between MOS and six of the most popular objective measures obtained from this experiment.

Performance of the DFTKCN in presence of car and train noise is evaluated using Itakura-Saito distance (ISD), Log-Likelihood ratio (LLR) [9] and Perceptual Evaluation of Speech Quality (PESQ) scores. One hundred sentences spoken by 20 speakers (10 Females and 10 Males) are randomly selected from WSJ database and contaminated by train and car noise at different noise levels. These noisy signals are then de-noised using PSS, MMSE, DFTKUN and DFTKCN methods and their distortion measures are obtained. The averaged results of the distortion measures are summarized in Table 4.

5.3. Discussion

Informal listening tests and comparisons of the quality of the output of the DFTKUN and DFTKCN methods with the MMSE log-STSA method reveal some major differences. The level of residual noise of DFT-Kalman methods is much less than that of MMSE. While DFTKUN slightly distorts the low energy portions of speech signal spectra as a result of the convergence of signal to small values. Due to this effect, at lower SNRs, the harmonics of the speech are well restored while the non-harmonic portions of the speech spectrum are relatively suppressed. This effect is mitigated in DFTKCN, while maintaining a similar or lower level of residual noise. Moreover, DFTKCN results in much less echo level than DFTKUN method producing a more natural-sounding speech signal. While the nature of the residual noise in spectral subtraction is musical (short bursts of narrowband energy), the residual noise of DFT-Kalman methods seems to have the same perceptual characteristic of the original noise.

Table 4: PESQ, LLR and ISD scores for various noise levels and types, obtained using different de-noising methods

Measure	Method	Car Noise SNR (dB)				Train Noise SNR (dB)			
		-5	0	5	10	-5	0	5	10
PESQ	DFTKUN	2.41	2.80	3.13	3.43	1.81	2.22	2.62	2.98
	DFTKCN	2.51	2.90	3.20	3.49	1.90	2.30	2.69	3.05
	MMSE	2.39	2.75	3.10	3.38	1.78	2.20	2.58	2.89
	PSS	2.44	2.79	3.08	3.28	1.65	2.12	2.51	2.84
LLR	DFTKUN	1.59	1.23	0.95	0.75	2.22	1.74	1.35	1.03
	DFTKCN	1.52	1.18	0.90	0.68	2.09	1.68	1.31	1.00
	MMSE	1.60	1.26	1.01	0.91	2.53	2.07	1.61	1.19
	PSS	1.59	1.25	1.01	0.87	2.64	2.17	1.67	1.23
ISD	DFTKUN	1.08	0.78	0.58	0.44	2.63	1.82	1.20	0.81
	DFTKCN	1.15	0.85	0.64	0.49	2.56	1.75	1.17	0.80
	MMSE	1.27	0.93	0.71	0.54	3.07	2.33	1.61	1.08
	PSS	1.41	1.04	0.77	0.59	3.43	2.71	1.89	1.19

6. CONCLUSION

A method is proposed for the enhancement of speech signals corrupted with background noise. The overall performance of the proposed method is shown to outperform MMSE log-STSA estimator and parametric spectral subtraction. Listening tests show that the residual noise of DFT-Kalman methods is not composed of annoying narrowband noise bursts, 'musical tones'. Informal experiments show that if the AR model of the DFT trajectories of clean speech are provided to the system (even in the case of using averaged models for the noise obtained from noise-only periods), the DFTKCN results in exceptionally superb quality of the de-noised speech. This suggests that the use of more sophisticated methods for estimation of the speech AR models is expected to result in further gain in the performance of the DFT-Kalman methods. The application of Expectation-Maximization (EM) methods for this purpose is being studied [10].

7. REFERENCES

[1] Ephraim, Y., Malah, D., "*Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator*", IEEE Trans. ASSP on Acoustics, Speech, and Signal Processing, vol. -32, no. 6, pp. 1109-1121, Dec. 1984.

[2] Sim, B., Tong, Y., Chang, J., Tan, C., "*A Parametric Formulation of the Generalized Spectral Subtraction Method*", IEEE Trans. on Speech and Audio Processing, vol. 6, No. 4, July 1998, pp. 328-337.

[3] Martin, R., *"Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors*", IEEE ICASSP'02, Orlando, Florida, May 2002.

[4] Cohen, I., "*On the Decision-Directed Estimation Approach of Ephraim and Malah*", ICASSP 04, Montreal, Canada, 17-21 May 2004, pp. I-293-296

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

[6] Kullback, S., Leibler, R.A., "*On information and sufficiency*", Ann. Math. Stat., vol. 22, pp. 79-86, 1951

[7] Brillinger, D.R., "*Time Series: Data Analysis and Theory*", Holden-Day, 1981

[8] E. Zavarehei, S. Vaseghi, "*Speech Enhancement In Temporal DFT Trajectories Using Kalman Filters*", Interspeech 2005, pp. 2077-2080

[9] Hansen, J., Pellom, B., "*An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms*", proc. of ICSLP 1998, Sydney

[10] Gannot, S., Burshtein, D., Weinstein, E., "*Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms*", IEEE Trans. on Speech and Audio Proc., vol. 6, no. 4, pp. 373-385, Jul. 1998