

SPEECH ENHANCEMENT IN TEMPORAL DFT TRAJECTORIES USING KALMAN FILTERS

Esfandiar Zavarehei, Saeed Vaseghi

Department of Electronic and Computer Engineering,
Brunel University, London, UK

{esfandiar.zavarehei, saeed.vaseghi}@brunel.ac.uk

Abstract

In this paper a time-frequency estimator for enhancement of noisy speech signals in the DFT domain is introduced. This estimator is based on modelling and filtering the temporal trajectories of the DFT components of noisy speech signal using Kalman filters. The time-varying trajectory of the DFT components of speech is modelled by a low order autoregressive process incorporated in the state equation of Kalman filter. The performance of the proposed method for the enhancement of noisy speech is evaluated and compared with MMSE log-STSA estimator and parametric spectral subtraction. Evaluation results show that the incorporation of temporal information through Kalman filters results in reduced residual noise and improved 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 [1-3]. 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 at each frequency. The phase distortion is assumed to be inaudible.

The various approaches proposed for estimation of the STSA of speech differ in three main aspects: (i) the use of different error functions such as STSA estimation error, log-STSA estimation error and short-time (ST) power spectrum estimation error [1][3], (ii) the use of different methods for estimation of speech statistics such as decision-directed method and non-causal estimation of a priori SNR [4] and (iii) the use of different probability distributions for speech spectral components such as Gaussian and Laplacian distributions [5].

An alternative to estimation of STSA is the estimation of the real and imaginary components of the ST-DFT of the clean speech. The MMSE estimation of ST-DFT components with Gaussian priors, leads to the well-known Wiener filter solution [6] while MMSE estimation of 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 modelling the real and imaginary components of ST-DFT of speech [6][7]. The use of ST-DFT with Gaussian priors lends itself to application of Kalman filter for modelling the temporal trajectory of speech.

Estimation of the statistics of clean speech spectrum, and the SNRs at different frequencies, are central to speech enhancement. Ephraim [1] employs a decision-directed

method for estimation of SNRs and tracking of speech statistics. The Markovian form of the decision-directed method [1] and the relatively high weight (0.95 to 0.99) given to the previous estimates of speech amplitude, show the importance of maintaining the temporal continuity of the speech spectrum. This, however, conflicts with the simplifying assumption of independent identical distribution of successive speech frames made in most MMSE estimators [1][4][6][7].

The modelling and utilisation of the time-varying trajectory of speech spectrum is the main focus of this paper. Our experiments show that there is a high level of dependency (correlation) between spectral samples of successive speech frames. In this paper, the temporal trajectory of speech spectrum is used in a more rigorous mathematical framework for a more reliable estimation of speech spectra. A set of linear prediction (autoregressive) models are incorporated in Kalman filters for adaptive estimation and modelling of the temporal trajectories of the ST-DFT components of the speech signals.

The rest of this paper is organized as follows. Section 2 discusses the modelling of the dependency of the samples of the temporal trajectories of ST-DFT components. In Section 3 the Kalman estimator of ST-DFT trajectories is introduced. In Section 4 the empirical issues of the new estimator are discussed and the evaluation results are compared with other methods of speech enhancement. The paper concludes in Section 5.

2. Modelling DFT Trajectories

In this section the temporal dependency and predictability of the trajectory of the ST-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 highly dependent on the frequency channel and the phoneme under study. Figure 1 illustrates the distribution of DFT components of channel 4 (120 Hz) for phoneme //l/. The data is obtained from 130 sentences spoken by a male speaker selected randomly from the Wall Street Journal database. It is evident from Figure 1 that the peak of the histogram is modelled better with a Laplacian distribution while the sides tend to fit a Gaussian distribution. This combination is observed repeatedly in different phonemes and different channels. The average symmetric Kullback-Leibler distance [8] between histograms and Laplacian distributions was calculated as 0.6 and between histograms and Gaussian distributions was calculated 0.8. This shows that on average a Laplacian distribution models the distribution of DFTs better than a Gaussian distribution. However, as often, a compromise between the deficiency of the model and the mathematical tractability of modelling the

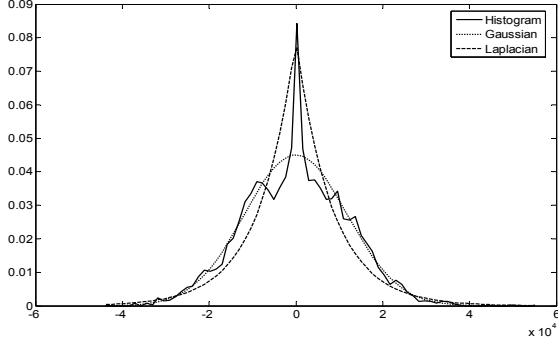


Figure 1: Normalized Histogram of DFT components channel 4 (120 Hz), Phoneme /l/, total number of data samples 4620. Window size: 25ms, Overlap: 15ms

temporal trajectories of DFT suggests the use of Gaussian distribution.

The real part of the ST-DFT of clean speech, $S_r(n)$, can be modelled using an AR process:

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

where $S_r(n)$ is the real part of the ST-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 distribution for ST-DFT components, the MMSE linear predictor (LP) coefficients of equation 1 can be obtained using Yule-Walker equation:

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

where $\mathbf{R}_{s_r}(n)$ and $\mathbf{r}_{s_r}(n)$ are the autocorrelation matrix and vector of $S_r(n)$ respectively and $\mathbf{a}(n)$ is the AR coefficient vector at frame n . The same equation stands for imaginary parts of the ST-DFT. The overlap size and LP order should be carefully chosen to comply with the stationarity assumption of equation 1, i.e. 20-40 ms.

Figure 2 illustrates the spectrogram of a sample signal reconstructed by predicting its temporal ST-DFT trajectories from their past values. The LP model order used for this

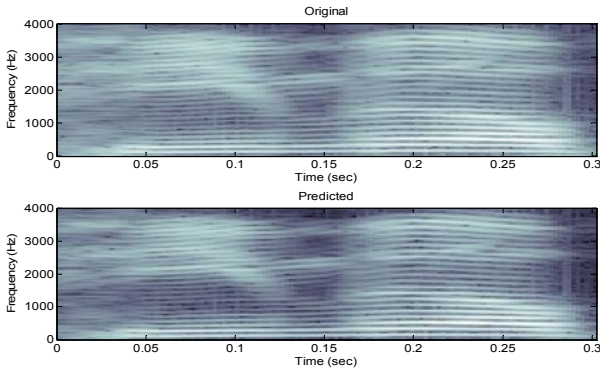


Figure 2: spectrogram of (top) original speech and (bottom) ST-DFT-predicted speech.

prediction is equal to $N=4$, which for a frame shift size of 5ms results in prediction of each sample from the previous samples in a window of 20ms. The coefficients of the LP models were obtained from 8 samples (40 ms) of each trajectory. After prediction of ST-DFT trajectories of the clean speech, the frames are overlap-added to reconstruct the ST-DFT-predicted speech signal. The differences that are observed from the spectrogram are a slight suppression of low energy portions and relative strengthening of high energy harmonics. No audible difference is observed between the actual clean signal and the LP-predicted signal.

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:

$$x(t) = s(t) + d(t) \quad (3)$$

For each frequency channel equation (3) is rewritten in ST-DFT domain as:

$$X_r(n) + jX_i(n) = (S_r(n) + D_r(n)) + j(S_i(n) + D_i(n)) \quad (4)$$

where the subscripts r and i represent the real and imaginary parts of ST-DFT respectively. It is assumed the real and imaginary parts of the ST-DFT are independent and have Gaussian distributions. This is verified from a study of the scatter plots of the real and imaginary parts of the DFT coefficients of clean speech [6] [9].

The trajectory of each component in each frequency channel can be modelled using an autoregressive model as in equation 1. Assuming a Gaussian distribution for ST-DFT, $e_r(n)$ is a zero mean Gaussian random variable, with a variance of $\sigma_r^2(n)$, and orthogonal to all previous values of S_r . From equations 1 and 4 the real component of ST-DFT can be rewritten in canonical form:

$$\mathbf{S}_r(n) = \mathbf{F}_r(n) \mathbf{S}_r(n-1) + \mathbf{G} e_r(n) \quad (5)$$

$$X_r(n) = \mathbf{H} \mathbf{S}_r(n) + D_r(n) \quad (6)$$

where

$$\mathbf{S}_r(n) = [S_r(n-N+1) \dots S_r(n)]^T \quad (7)$$

$$\mathbf{F}_r(n) = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ a_N(n) & a_{N-1}(n) & a_{N-2}(n) & \dots & a_1(n) \end{bmatrix} \quad (8)$$

$$\mathbf{G}^T = [0 \dots 0 \ 1] = \mathbf{H} \quad (9)$$

and $\mathbf{S}_r(n)$ is the state vector. Using equation 5, given the previous estimate of the state vector and the AR model, the MMSE prediction of the state vector, $\hat{\mathbf{S}}_r^-(n)$, is obtained as:

$$\hat{\mathbf{S}}_r^-(n) = E\{\mathbf{S}_r(n) | \hat{\mathbf{S}}_r^-(n-1)\} = \mathbf{F}_r(n) \hat{\mathbf{S}}_r^-(n-1) \quad (10)$$

where $\hat{\mathbf{S}}_r(n-1)$ is the *estimate* of $\mathbf{S}_r(n-1)$. As $e_r(n)$ is orthogonal to $\hat{\mathbf{S}}_r(n-1)$, the *prediction* error covariance matrix is calculated as:

$$\mathbf{P}_r^-(n) = \mathbf{F}_r(n)\mathbf{P}_r(n-1)\mathbf{F}_r^T(n) + \mathbf{G}\mathbf{G}^T\sigma_r^2(n) \quad (11)$$

where $\mathbf{P}_r(n-1)$ is the state *estimation* error covariance matrix. Incorporating the innovation signal (i.e. the difference between the prediction of speech and the noisy speech) in the current noisy observation, the optimum estimate of state vector is calculated as:

$$\hat{\mathbf{S}}_r(n) = \hat{\mathbf{S}}_r^-(n) + \mathbf{K}_r(n)(X_r(n) - \mathbf{H}\hat{\mathbf{S}}_r^-(n)) \quad (12)$$

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

$$\mathbf{K}_r(n) = \mathbf{P}_r^-(n)\mathbf{H}^T[\mathbf{H}\mathbf{P}_r^-(n)\mathbf{H}^T + \nu_r^2(n)]^{-1} \quad (13)$$

and $\nu_r^2(n)$ is the variance of noise, $D_r(n)$. Note that equation (13) does not involve any matrix inversion as $\mathbf{H}\mathbf{P}_r^-(n)\mathbf{H}^T$ is a scalar. The *estimation* error covariance of this estimate which is required for the next step is obtained as:

$$\mathbf{P}_r(n) = [\mathbf{I} - \mathbf{K}_r(n)\mathbf{H}]\mathbf{P}_r^-(n) \quad (14)$$

The same equations hold for the imaginary component of all frequency channels. The first (DC) and last frequency channels, however, have only the real component as imposed by Fourier transform equations.

4. Implementation of DFT-Kalman Filters

4.1 DFT-Kalman Filters for Speech Enhancement

A set of Kalman filters are implemented in parallel for modelling and estimation of the ST-DFT trajectories. As the clean speech signal is not available, the estimates of AR models for each track are obtained from the noisy/restored data. In this work the estimates of AR model of each channel at frame n was obtained from the previous $L=8$ estimates of the restored speech ST-DFT trajectory of that channel. Modifying equation 2 the estimate of $\mathbf{a}(n)$ is obtained as:

$$\hat{\mathbf{a}}(n) = (\mathbf{R}_{s_r}(n-1))^{-1} \times \mathbf{r}_{s_r}(n-1) \quad (15)$$

An implementation issue arises from the feedback of restored speech for calculation of AR parameters using equation 15. During long (typically >200 ms) noise-only periods, where the variance of the noisy signal is equal to that of noise, the recursive solution given by equations (11) and (13-15), results in convergence of the output of equation (12) towards zero which consequently decreases the variance of prediction error, $\sigma_r^2(n)$, towards zero. In order to prevent the consequent zeroing of speech following a long period of speech inactivity the value of $\sigma_r^2(n)$ needs to be revived from zero at the beginning of speech active periods. This is achieved by ensuring that values of $\sigma_r^2(n)$ will not be less than a factor α of the noisy signal, using the following equation:

$$\hat{\sigma}_r^2(n) = \max\left(\sigma_r^2(n), \alpha^2 |X^2(n)|\right) \quad (16)$$

which limits the prediction error variance to a small portion of the instantaneous power spectrum of noisy speech. Equation

Table 1: DFT-Kalman Speech Enhancement Algorithm

Initializations: Obtain initial mean and variance of noise spectrum and initialize $\mathbf{P}(n)$

For each speech frame:

For each frequency channel:

Compute AR model coefficients, \mathbf{a} (15)

Compute AR model excitation variance (16)

Predict speech $\hat{\mathbf{S}}^-(n)$ (10)

Compute Kalman gain, $\mathbf{K}(n)$ (11,13)

Estimate speech $\hat{\mathbf{S}}(n)$ (12)

Estimate error covariance matrix $\mathbf{P}(n)$ (14)

(16) implies that the ST-DFT trajectories can be predicted with a limited precision, i.e. the prediction error variance cannot be smaller than a threshold proportional to the variance of noisy speech. Very small values for α proved to be sufficient for reviving the converged trajectories of $\sigma_r^2(n)$ and the signal at the beginning of speech activity (e.g. $\alpha=0.07$). The steps of the DFT-Kalman spectral enhancement algorithm are summarized in Table 1.

4.2 Objective and Subjective Evaluations

The evaluation of the performance of DFT-Kalman filter 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 Wall Street Journal speech database. The noisy signals are segmented using 25ms hamming windows with an overlap size of 20ms. The car noise signal is recorded by our colleagues in a BMW at 70 Mph in a rainy day and the train noise is recorded in a moving train.

4.2.1 Mean Opinion Score (MOS)

A set of five sample sentences are drawn from WSJ database (3 female, 2 male) and contaminated by car noise, train noise and white Gaussian noise (WGN) at two different SNRs, 0dB and 10dB. The resulting 30 noisy speech sentences are then de-noised using three different methods: (i) parametric spectral subtraction (PSS) [2], (ii) MMSE log-STSA [3] and (iii) the proposed DFT-Kalman (DFTK) method. Ten 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. DFTK proved to perform consistently better in terms of MOS. It was observed that the performance of DFTK is especially preferred in highly non-stationary train noise. As usual, the extent of the validity of these results is limited by the method used for evaluation, the number of sample signals and number of listeners.

4.2.2 Objective Evaluation

From a number of different speech quality and distortion measures applied to the restored sample speech sentences of section 4.2.1, 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 purpose. Table 3 summarizes the correlation coefficients between MOS and six of the most popular objective measures obtained from this experiment.

Table 2: Mean opinion score results

SNR	Noise	DFTK	MMSE	PSS
0dB	Car	3.7	3.5	3.4
	Train	2.7	2.0	2.1
	WGN	2.0	1.6	1.4
10dB	Car	4.5	4.6	4.2
	Train	3.7	3.7	3.5
	WGN	3.3	3.2	2.4

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

With regards to the results of Table 3, the performance of the DFT-Kalman method in presence of car and train noise is evaluated using Itakura-Saito distance (ISD), Log-Likelihood ratio (LLR) [10] and PESQ (Perceptual Evaluation of Speech Quality) 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 and DFTK methods and their distortion measures are obtained. Table 4 summarizes the averaged results of the distortion measures. It is evident that DFT-Kalman performs consistently better than the other methods.

4.2.3. Discussion

Informal listening tests and comparisons of the quality of the output of the DFT-Kalman method with the MMSE log-STSA method reveal three major differences. First, the level of residual noise of DFT-Kalman is much less than that of MMSE. Second, DFT-Kalman slightly distorts the low energy portions of speech signal spectra as a result of the convergence of signal to small values. These distortions are not significant but with careful listening are audible. Third, some echo is audible if relatively high order LP models ($N > 8$) are used or the shift size is increased. If equation (16), which sets a limit on prediction error variance, is applied regardless of speech activity, a small amount of noise will remain in the signal but the speech distortion and echo will not be audible anymore. Moreover, from comparison of the DFT-Kalman method and parametric spectral subtraction it is revealed that while the nature of the residual noise in spectral subtraction is musical (short bursts of narrowband energy) this effect is not observed in DFT-Kalman.

5. 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. It is observed that incorporating the information on temporal evolution of the trajectories through Kalman filter and AR models results in better performance of de-noising procedure in terms of objective and subjective quality evaluations. Moreover, informal listening tests show that the residual noise of DFT-Kalman method is not composed of annoying narrowband noise bursts, ‘musical tones’. The applicability of DFT-Kalman as a de-noising algorithm for robust speech recognition in noisy environments is being investigated.

Table 4: Itakura-Saito Distance (ISD), Log-likelihood Ratio (LLR) and PESQ scores for various noise levels and types, obtained using different de-noising methods

Measure	Noise Type	Method	Input SNR (dB)			
			-5	0	5	10
ISD	Car	MMSE	1.48	1.18	0.93	0.73
		PSS	1.49	1.12	0.84	0.67
		DFTK	1.06	0.77	0.61	0.49
	Train	MMSE	2.62	2.07	1.55	1.19
		PSS	2.84	2.08	1.49	1.06
		DFTK	2.16	1.54	1.11	0.8
LLR	Car	MMSE	1.70	1.41	1.14	0.91
		PSS	1.70	1.34	1.07	0.88
		DFTK	1.55	1.18	0.95	0.77
	Train	MMSE	2.43	2.00	1.63	1.31
		PSS	2.47	1.95	1.55	1.21
		DFTK	2.21	1.74	1.35	1.05
PESQ	Car	MMSE	2.35	2.71	3.03	3.29
		PSS	2.38	2.74	3.07	3.3
		DFTK	2.39	2.76	3.13	3.45
	Train	MMSE	1.77	2.22	2.57	2.85
		PSS	1.75	2.19	2.59	2.88
		DFTK	1.81	2.22	2.62	3.01

6. Reference

- [1] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.
- [2] B. Sim, Y. Tong, J. Chang, C. Tan, “A Parametric Formulation of the Generalized Spectral Subtraction Method”, IEEE Transactions on Speech and Audio Processing, vol. 6, No. 4, July 1998, pp. 328-337.
- [3] Y. Ephraim and D. Malah, “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.
- [4] I. Cohen, “On the Decision-Directed Estimation Approach of Ephraim and Malah”, Proc. 29th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2004, Montreal, Canada, 17-21 May 2004, pp. I-293-296.
- [5] B. Chen, P. Loizou, “Speech Enhancement Using a MMSE Short Time Spectral Amplitude Estimator with Laplacian Speech Modeling”, to be published in ICASSP2005
- [6] R. Martin, “Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors”, IEEE ICASSP’02, Orlando, Florida, May 2002.
- [7] R. Martin, Breithaupt, C., “Speech Enhancement in the DFT Domain Using Laplacian Speech Priors”, Proc. Int. Workshop Acoustic Echo and Noise Control (IWAENC), 2003, pp.87-90
- [8] S. Kullback, R.A. Leibler, “On information and sufficiency”, Ann. Math. Stat., vol. 22, pp. 79-86, 1951
- [9] D.R. Brillinger, “Time Series: Data Analysis and Theory”, Holden-Day, 1981
- [10] J. Hansen, B. Pellom, “An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms”, proc. of ICSLP’98, Sydney, 1998