NOISE POWER SPECTRUM ESTIMATION FOR SPEECH ENHANCEMENT USING AN AUTOREGRESSIVE MODEL FOR SPEECH POWER SPECTRUM DYNAMICS

Ivo Batina, Jesper Jensen and Richard Heusdens

Information and Communication Theory Group,
Delft University of Technology,
2628 CD Delft, The Netherlands,
email:{i.batina,j.jensen,r.heusdens}@ewi.tudelft.nl

ABSTRACT

In this paper we propose a method for estimating the non-stationary noise power spectral density (PSD) given a noisy speech signal. The method is based on an autoregressive (AR) model of the speech PSD dynamics combined with a Kalman filtering based noise PSD estimation technique. Objective and subjective performance evaluations show that the speech enhancement scheme utilizing the proposed noise PSD estimation technique achieves significant improvements over a system using a stationary noise estimate as well as compared to a system that uses a noise tracker developed in our previous work.

1. INTRODUCTION

The class of speech enhancement techniques based on short-time spectral amplitude (STSA) estimation (see [1, 2, 3]) have proved to be of particular practical interest due to their low complexity and relatively good performance. As most single-channel speech enhancement (SE) methods, STSA based techniques require a power spectral estimate of the noise process in order to extract a clean speech signal estimate from a noisy realization. As any SE scheme, the performance of STSA based techniques is much affected by the capability to track variations in the statistics of the noise [4], particularly under low signal-to-noise ratio (SNR) conditions and non-stationary noise environments. In [4] a recursive scheme for noise estimation, commonly known as the minimum Statistics (MS) method, is designed to be combined with STSA speech enhancement schemes. The method is based on tracking the noisy speech spectral minima without any distinction between speech activity and speech pause, enabling the algorithm to update the noise estimate even in the regions where speech is present. A similar method is described in [5], where the response of the noise estimator to the rise of the noise level is improved by periodogram smoothing across both time and frequency and speech presence probability estimation.

In [6], we proposed a method for noise power spectral density (PSD) estimate that is based on the application of the Kalman filtering technique in the STSA context. Objective and subjective performance evaluations [6] showed that the proposed scheme exhibits a good noise tracking performance and that it achieves improvement in the quality of the enhanced speech as compared to the case where the noise PSD estimate remains invariant across time. Listening test results indicated a statistically significant improvement in the quality of enhanced speech compared to the fixed noise PSD estimate case.

This research was partly supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs.

The Kalman filtering based noise PSD scheme in [6] is based on a simple, low order model of the speech power spectrum. In this paper we look into the speech PSD modelling problem and derive an AR model of the speech PSD dynamics. The model is then used in the Kalman filtering based noise PSD estimator [6]. Both subjective and objective evaluation results show improvement of the performance with the model of the speech PSD dynamics presented in this paper.

2. STOCHASTIC MODEL OF THE NOISY SPEECH POWER SPECTRUM

To derive a stochastic model of the noisy speech power spectrum, we assume that the noise is additive. Therefore, the short-time Fourier transform (STFT) of the noisy speech signal can be written as

$$Y(k,l) = S(k,l) + N(k,l) \tag{1}$$

where $Y(k,l)$, $S(k,l)$, $N(k,l)$ denote STFT coefficients of the noisy speech signal, clean speech and the noise, respectively, k denotes the frequency bin index and l represents frame index. Furthermore, we assume that speech and noise are uncorrelated random processes and that the STFT coefficients are Gaussian complex variables (see e.g. [2, 3]). It can then be shown that the magnitude square of the noisy speech STFT coefficients are exponentially distributed random variables for all k and l with a probability density function given by

$$f_{|Y(k,l)|^2}(x) = \frac{1}{\lambda_s(k,l) + \lambda_n(k,l)} \exp\Big\{ - \frac{x}{\lambda_s(k,l) + \lambda_n(k,l)} \Big\} \tag{2}$$

with $x \geq 0$. Variances of the speech and the noise STFT coefficients in (2) are denoted $\lambda_s(k,l) = \mathbb{E}\{|S(k,l)|^2\}$ and $\lambda_n(k,l) = \mathbb{E}\{|N(k,l)|^2\}$, respectively.

Next, define the stochastic process

$$y(k,l) = \big(\lambda_s(k,l) + \lambda_n(k,l)\big) e(k,l) \tag{3}$$

where $e(k,l)$ is an exponentially distributed random variable with mean and variance equal to 1. It is easy to verify that the probability density function (pdf) of $y(k,l)$ is identical to that of $|Y(k,l)|^2$ given in (2).

Since the noise and speech processes are generally non-stationary, their respective PSD's change across time. Consequently, the PSD of the noisy signal is also time-varying. In the following we set up a model to represent these variations of the speech PSD.

To derive the model, we assume that the time-series consisting of the squared magnitude of the clean speech STFT coefficients at a

particular frequency bin index k can be modelled by a linear, nonzero mean, time varying AR model as

$$s(k,l) - \sum_{i=1}^{N_k} a_i(k,l)s(k,l-i) = e_s(k,l)w_s(k,l) + \delta(k,l) \quad (4)$$

with $s(k,l) = |S(k,l)|^2$, where $\delta(k,l)$ is the mean of the model, $w_s(k,l)$ is drawn from a white Gaussian noise process with zero mean and variance one and $e_s(k,l)$ is a positive, real valued parameter.

Because $\lambda_s(k,l)$ can not be observed from the clean speech periodogram, we follow the strategy of [4] and use a recursive periodogram smoothing technique to obtain an estimate of the clean speech power spectrum

$$\lambda_s(k,l) = \alpha\lambda_s(k,l-1) + (1-\alpha)s(k,l). \quad (5)$$

After computing the mean $\delta(k,l)$ by taking expectation on both sides of (4) and substituting (4) in (5) we obtain

$$s(k,l) = \sum_{i=1}^{N_k} \frac{a_i(k,l)}{1-(1-\alpha)\Phi(k,l)} s(k,l-i) + \\ + \frac{e_s(k,l)}{1-(1-\alpha)\Phi(k,l)} w_s(k,l) + \frac{\alpha\Phi(k,l)}{1-(1-\alpha)\Phi(k,l)} \lambda_s(k,l-1) \quad (6)$$

where

$$\Phi(k,l) = 1 - a_1(k,l) - a_2(k,l) \ldots - a_{N_k}(k,l). \quad (7)$$

We model the noise PSD dynamics using a linear, time varying AR model. For simplicity, we use a first order model from [6]

$$\lambda_n(k,l+1) - b_1(k,l)\lambda_n(k,l) = e_n(k,l)w_n(k,l) \quad (8)$$

where $w_n(k,l)$ is drawn from a white Gaussian noise process with zero mean and variance one and $e_n(k,l)$ is a positive, real valued parameter. It is straightforward to generalize the noise model for the higher orders. In order to use (4), (5) and (8) in the noise PSD estimator based on Kalman filtering [6] we need to rewrite these equations in the state space form. To this aim, we introduce state variables $r(k,l) = \alpha\lambda_s(k,l-1)$, $\lambda_n(k,l)$ and

$$q(k,l) = \frac{a_1(k,l)}{1-(1-\alpha)\Phi(k,l)} s(k,l-1).$$

We can rewrite (6) and (8) in state space form as

$$x(k,l+1) = A(k,l)x(k,l) + E(k,l)w(k,l) \quad (9)$$

where

$$x(k,l) = \begin{bmatrix} q(k,l) & \cdots & q(k,l-N_k+1) & r(k,l) & \lambda_n(k,l) \end{bmatrix}^T$$

is the state vector,

$$A(k,l) = \begin{bmatrix} \beta_1(k,l) & \cdots & \beta_{N_k}(k,l) & \beta_\lambda(k,l) & 0 \\ 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ \gamma_1(k,l) & \cdots & \gamma_{N_k}(k,l) & \alpha + (1-\alpha)\beta_\lambda(k,l) & 0 \\ 0 & \cdots & 0 & 0 & b_1(k,l) \end{bmatrix},$$

$$E(k,l) = \begin{bmatrix} \frac{\beta_1(k,l)e_s(k,l)}{1-(1-\alpha)\Phi(k,l)} & 0 & \cdots & 0 & \frac{\alpha(1-\alpha)e_s(k,l)}{1-(1-\alpha)\Phi(k,l)} & e_n(k,l) \end{bmatrix}^T,$$

$$\beta_i(k) = \frac{a_i(k,l)}{1-(1-\alpha)\Phi(k,l)}, \quad \beta_\lambda(k,l) = \frac{\Phi(k,l)\alpha}{1-(1-\alpha)\Phi(k,l)},$$

$$\gamma_i(k,l) = \alpha(1-\alpha)\frac{\beta_i(k,l)}{\beta_1(k,l)}$$

and where $w(k,l)$ is drawn from a white Gaussian noise process with zero mean and variance one. Finally, we write (3) in matrix form as

$$y(k,l) = Cx(k,l)e(k,l) \quad (10)$$

where

$$C = \begin{bmatrix} 0 & \cdots & 0 & \frac{1}{\alpha} & 1 \end{bmatrix}.$$

Since the variances of the speech and the noise processes and the squared magnitude of the STFT coefficients must be larger than zero, the entries in the state vector $x(k,l)$ have to satisfy the following constraint

$$q(k,l) \geq 0, \quad r(k,l) \geq 0 \quad \text{and} \quad \lambda_n(k,l) \geq 0 \quad (11)$$

for all k, l. Equations (10) and (9) together with the constraint (11) form the state space model for the noisy speech power spectrum dynamics.

3. ESTIMATION OF THE MODEL PARAMETERS

To use the model presented in Section 2 we must determine the model order and the model parameters $a_i(k,l)$, $e_s(k,l)$ in (4) and $b_1(k,l)$, $e_n(k,l)$ in (8). We note that the state space model formulated in (9), (10) allows for time varying model parameters. Thus, the model parameters can be estimated from the noisy data on a frame-by-frame basis or recursively updated from past estimates in an adaptive model estimation structure. In this paper we choose a simpler approach where the model parameters are estimated in an off-line estimation procedure. At run-time the estimated model parameters remains constant across time but they vary across frequency. Our aim was to investigate a possibility to improve the performance of the scheme in [6] with the least added computational complexity. Note that by using time invariant coefficients of the model we consider average spectral behavior. Performance evaluation results reported in this paper show that taking into account average spectral behavior improves performance of the scheme in [6]. Further improvements that can be achieved by time varying model of the noisy speech PSD dynamics are topics for further research.

To estimate the parameters of the AR model (4), we consider a clean speech signal that consists of four different speech utterances. The utterances, two from male and two from female speakers are taken from the TIMIT database and downsampled to 8 kHz. We obtain the squared magnitude of the speech STFT coefficients $s(k,l) = |S(k,l)|^2$ by using the discrete Fourier transform of the signal frames extracted with a Hanning window of 256 samples and use of an inter frame overlap of 50%. These settings are the same as those of the STFT enhancement scheme used for the performance evaluation reported later. Then, we use the Levinson-Durbin algorithm as described in [7] to solve augmented Wiener-Hopf equation for forward linear prediction. Coefficient $e_n(k,l)$ is obtained from the variance of the linear prediction error.

In this paper, we set the clean speech AR model order $N_k = 2$ for all k. Experiments with a larger order model do not show improvement in the noise PSD tracking performance. Together with the smoother (5) and noise model (8) the state space model (9), (10)

that is used for Kalman filtering has the order of 4. We note here the AR model order does not need to be equal for all frequencies. Especially for the frequency bands that contain significant speech energy further improvement can be obtained by finding an optimal "model order distribution" over frequency. This is topic for further research.

4. KALMAN FILTERING BASED NOISE POWER SPECTRAL DENSITY ESTIMATOR

The model of the noise PSD dynamics given in Section 2 with parameters estimated as described in Section 3 is used in the Kalman filtering based noise PSD estimator [6]. The Kalman filtering structure for the noise PSD estimation is given by

$$\hat{x}(k,l+1) = A(k)\hat{x}(k,l) + K(k,l)\left(|Y(k,l)|^2 - C\hat{x}(k,l)\right) \quad (12)$$

with the Kalman gain given by

$$K(k,l) = A(k)Q_e(k,l)C^T\left(C\big(2Q_e(k,l) + \hat{Q}(k,l)\big)C^T\right)^{-1} \quad (13)$$

where

$$\begin{aligned} Q_e(k,l+1) = \big(A(k) - K(k,l)C\big)Q_e(k,l)\big(A(k) - K(k,l)C\big)^T + \\ K(k,l)CQ_e(k,l)C^TK^T(k,l) + \\ C\hat{Q}(k,l)C^T + EQ_w(k,l)E^T \end{aligned}$$

is the variance of the estimation error and

$$\begin{aligned} \hat{Q}(k,l+1) = \big(A(k) + K(k,l)C\big)\hat{Q}(k,l)\big(A(k) + K(k,l)C\big)^T + \\ 2K(k,l)CQ_e(k,l)C^TK^T(k,l) + EQ_w(k,l)E^T. \end{aligned}$$

Implementation of the Kalman filter (12) for the new model of the noisy speech power spectrum dynamics is straightforward.

5. PERFORMANCE EVALUATION

The performance evaluation of the Kalman filtering based noise tracking algorithm [6] with the improved model of the speech PSD dynamics (9), (10) consists of two parts. First we show the tracking capability of the algorithm for nonstationary white noise. Secondly, we compare the performance of the estimator [6] with the scheme proposed in this paper. Speech enhancement in the evaluations has been performed by using the log spectral amplitude (LSA) enhancement scheme [3]. We perform an objective as well as a subjective quality assessment of the enhanced speech samples. Speech utterances that we use in evaluation of the proposed estimator are different from the utterances that are used in the estimation of the AR models of the clean speech PSD dynamics.

To show the tracking capability of the proposed noise PSD estimator we degrade the speech signal with non-stationary white Gaussian noise. The SNR in the noisy speech signal is 30 dB for the first 3.75 seconds of the signal. After that, the noise level rises with the constant rate of 0.15 dB/frame up to 0 dB SNR where it stays constant for the remaining part of the signal. To ease the visualisation of the results, we adopted the procedure used in [4, 5] and compute the average noise PSD estimate across frequency for each frame. We emphasize that this frequency averaging is only done for presentation purposes. The proposed method does not exploit a priori knowledge that the noise source in this case is spectrally flat. From

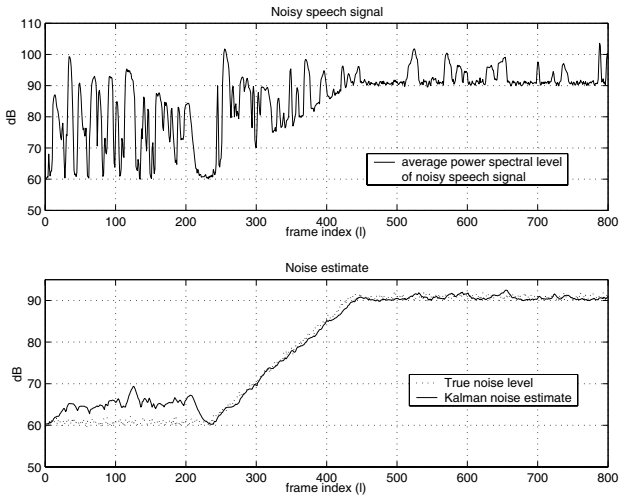

Fig. 1. The mean over frequency of the noisy speech sample power spectrum and the noise PSD estimates obtained with different noise estimators.

the Figure 1 we observe that the proposed estimator tracks variations in the noise level in the noise region even during the speech presence.

Next, we consider speech signals degraded by white, babble, factory and car noise, at various SNR levels. We compare the performance of the LSA enhancement scheme [3] for the noise PSD estimator [6] (KF in Table 1), the noise estimator proposed in this paper (ArKF in Table 1) and the case when no noise tracking is performed (NNT in Table 1). In this case we compute a Bartlett estimate of noise PSD in the noise only region preceding the speech signal and keep this value as the noise level estimate for the whole duration of the signal. We also give values of the various distortion measures for the noisy speech sample (NS in Table 1). For objective quality assessment we use the symmetric Itakura-Saito distortion measure (sym.I.S in Table 1) [8], segmental SNR measure (S.SNR in Table 1) and log likelihood ratio distortion measure (L.L.R. in Table 1) [9, 7]. The results of the objective quality assessment show that all considered distortion measures indicate improvement of the performance when the proposed noise PSD estimator (ArKF) is used in the enhancement scheme.

For subjective evaluation an OAB listening test was performed with nine participants, the authors not included. We compare the performance of the noise PSD estimator [6] and the proposed estimator. In this listening test we used babble, factory and car noise at 5 dB and 15 dB SNR. For each noise source and noise level we presented listeners two female and two male sentences. The listeners were presented first the noise free signal followed by the two different enhanced signal in randomized order, and this was repeated three times for each series. For speech signals corrupted with babble noise the proposed scheme was preferred above the noise tracker [6] in 79.15 % (15 dB) and 87.13 % (5 dB) of the cases, for factory noise in 87.5 % (15 dB) and 72.93 % (5 dB) cases and for car in 96.86 % (5 dB) and 100 % (15 dB).

6. CONCLUSIONS AND DISCUSSIONS

We presented a method for noise PSD tracking in noisy speech signals. The method is based on an autoregressive (AR) model of the

White noise												
	Input SNR 0 dB			Input SNR 5 dB			Input SNR 10 dB			Input SNR 15 dB		
	S.SNR	sym.I.S.	L.L.R	S.SNR	sym.I.S.	L.L.R.	S.SNR	sym.I.S.	L.L.R.	S.SNR	sym.I.S.	L.L.R
NS	-3.65	7191.3	1.66	-1.27	3177.1	1.55	2.03	2078.4	1.44	5.69	2485.0	1.39
NNT	-1.00	5633.4	1.53	1.26	1035.3	1.48	3.86	1734.2	1.34	5.91	1855.3	1.12
KF	-1.09	160.08	1.38	2.38	635.58	1.28	4.44	1035.4	1.25	6.73	334.4	1.07
ArKF	0.00	103.76	1.32	3.56	34.75	1.22	5.81	33.44	0.95	8.06	36.39	0.71
Babble noise												
	Input SNR 0 dB			Input SNR 5 dB			Input SNR 10 dB			Input SNR 15 dB		
	S.SNR	sym.I.S.	L.L.R	S.SNR	sym.I.S.	L.L.R.	S.SNR	sym.I.S.	L.L.R.	S.SNR	sym.I.S.	L.L.R
NS	-3.61	2363.2	1.87	-0.86	1337.8	1.76	2.42	999.39	1.67	6.20	379.79	0.63
NNT	-1.77	1838.9	1.33	0.66	1017.3	1.10	2.78	496.50	0.81	6.48	296.53	0.55
KF	-1.65	865.68	1.30	0.91	857.77	0.99	3.58	293.63	0.79	6.79	94.65	0.31
ArKF	-1.04	278.92	0.79	1.70	101.05	0.61	4.68	41.04	0.45	8.33	15.37	0.19
Factory noise												
	Input SNR 0 dB			Input SNR 5 dB			Input SNR 10 dB			Input SNR 15 dB		
	S.SNR	sym.I.S.	L.L.R	S.SNR	sym.I.S.	L.L.R.	S.SNR	sym.I.S.	L.L.R.	S.SNR	sym.I.S.	L.L.R
NS	-3.57	1939	1.36	-0.86	1146.7	2.25	2.34	3831.2	0.92	6.03	1513	0.74
NNT	-1.89	391.23	1.31	-0.03	1019.2	1.79	3.02	812.91	0.89	6.52	341.3	0.58
KF	-1.03	203.03	1.30	1.09	712.83	1.11	3.91	246.96	0.89	6.82	82.24	0.51
ArKF	0.01	94.47	0.87	2.39	60.26	0.68	5.29	45.35	0.50	8.24	40.98	0.35
Car noise												
	Input SNR 0 dB			Input SNR 5 dB			Input SNR 10 dB			Input SNR 15 dB		
	S.SNR	sym.I.S.	L.L.R	S.SNR	sym.I.S.	L.L.R.	S.SNR	sym.I.S.	L.L.R.	S.SNR	sym.I.S.	L.L.R
NS	-2.96	4079.3	0.51	-0.05	4777.3	0.41	3.36	5067.3	0.34	7.19	5892.7	0.25
NNT	1.54	3319.6	0.48	1.87	399.34	0.37	5.98	439.23	0.26	9.01	574.91	0.17
KF	2.32	379.22	0.49	5.17	134.08	0.33	7.29	44.73	0.22	9.76	14.57	0.15
ArKF	6.95	31.9	0.38	11.06	11.92	0.29	14.11	4.84	0.21	15.37	2.67	0.15

Table 1: Segmental SNR (S.SNR), symmetric Itakura-Saito (sym.I.S) and log likelihood ratio (L.L.R) distortion measure for white, babble, factory and car noise at various SNR levels using Kalman filtering based noise PSD estimators with different model of the noisy speech power spectrum

speech PSD dynamics combined with a Kalman filtering based noise power spectral density estimation technique [6]. Although we apply the method in the LSA based speech enhancement context, the method can be useful in other speech enhancement system that requires a noise power spectral estimate, e.g., codebook-driven methods and subspace based approaches.

We perform objective and subjective evaluation of the proposed method. As results presented in Section 5 show, the proposed noise PSD tracking method exhibits good noise tracking capabilities and evaluation experiments showed preference over the case when there is no noise tracking and the previously developed method in [6].

Further development of the model is to consider the time-varying model parameters case. We expect that the introduction of time-varying model parameters will lead to improved noise PSD estimates at the cost of higher computational load.

7. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwith compression of noisy speech," *Proc. of the IEEE*, vol. 67, no. 12, pp. 1586–1604, December 1979.

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

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[5] I. Cohen, "Noise spectrum estimation in adverse enviroments: Improved minima controller recursive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[6] I. Batina, J. Jensen, and R. Heusdens, "Kalman filtering based noise power spectral density estimation for speech enhancement," in *Proc. of the 13th European Signal Processing Conference*, September 2005.

[7] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, Macmillan, NewYork, 1993.

[8] A. H. Gray, Jr. and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-24, no. 5, pp. 380–391, October 1976.

[9] T. P. Barnwell S. R. Quackenbush and M. A. Clements, *Objective Measures of Speech Quality*, Advanced Reference Series. Prentice Hall, 1988.