SPEECH ENHANCEMENT USING AN MMSE SPECTRAL AMPLITUDE ESTIMATOR BASED ON A MODULATION DOMAIN KALMAN FILTER WITH A GAMMA PRIOR

Yu Wang[⋆] *Mike Brookes*[†]

[⋆]Engineering Department, University of Cambridge, United Kingdom
[†]Department of Electrical and Electronic Engineering, Imperial Collge London, United Kingdom
yw396@cam.ac.uk mike.brookes@imperial.ac.uk

ABSTRACT

In this paper, we propose a minimum mean square error spectral estimator for clean speech spectral amplitudes that uses a Kalman filter to model the temporal dynamics of the spectral amplitudes in the modulation domain. Using a two-parameter Gamma distribution to model the prior distribution of the speech spectral amplitudes, we derive closed form expressions for the posterior mean and variance of the spectral amplitudes as well as for the associated update step of the Kalman filter. The performance of the proposed algorithm is evaluated on the TIMIT core test set using the perceptual evaluation of speech quality (PESQ) measure and segmental SNR measure and is shown to give a consistent improvement over a wide range of SNRs when compared to competitive algorithms.

Index Terms— speech enhancement, modulation domain Kalman filter, minimum mean-square error (MMSE) estimator

1. INTRODUCTION

Over several decades, numerous speech enhancement algorithms have been proposed. Among the most popular are those such as [1, 2, 3] which apply a variable gain in the short time Fourier transform (STFT) domain to estimate the spectral amplitudes of the clean speech. Although these STFT-domain enhancement algorithms often improve the signal-to-noise ratio (SNR) dramatically, the temporal dynamics of the speech spectral amplitudes are not incorporated into the derivation of the estimator. There is evidence, however, that significant information in speech is carried by the modulation of spectral envelopes in addition to the envelopes themselves [4, 5]. Spectral modulation-domain processing has been used in speech recognition [6, 7], in speech intelligibility metrics [8, 9] and in speech enhancement [10, 11, 12]. In one such enhancement algorithm [12], the temporal envelope of the amplitude spectrum of the noisy speech is processed separately in each subband by a Kalman filter (KF) in order to obtain the spectral amplitudes of the enhanced speech. This modulation-domain KF combines the estimated dynamics of the speech spectral amplitudes with the observed noisy speech amplitudes to give an minimum mean square error (MMSE) estimate of the amplitude spectrum of the clean speech, under the assumption that the spectral amplitudes of both the clean speech and the noise are Gaussian distributed.

In this paper, we propose an MMSE spectral amplitude estimator under the assumption that the speech amplitudes follow a generalized Gamma distribution [13]. The advantage of the proposed estimator over previously proposed spectral amplitude estimators [2, 13, 14] is that it incorporates temporal continuity into the MMSE estimator by the use of the KF and that it uses a Gamma prior which is a more appropriate model for the speech spectral amplitudes than a Gaussian prior [11].

Yu Wang was a PhD student at Imperial College London during the course of this work.

2. SIGNAL MODEL AND KALMAN FILTER

We assume an additive model in the STFT domain in which, for frequency bin k of frame n,

$$Y_{n,k} = X_{n,k} + W_{n,k} \tag{1}$$

where X and W denote the complex-valued STFT coefficients of the clean speech and the noise respectively. Since each frequency bin is processed independently within our algorithm, we omit the frequency index, k, in the remainder of this paper. We denote the spectral amplitudes as: $|X_n| = A_n$, $|Y_n| = R_n$ and $|W_n| = N_n$. The prediction model we assume for the clean speech spectral amplitudes is

$$\mathbf{a}_n = \mathbf{F}_{n-1}\mathbf{a}_{n-1} + \mathbf{v}_n \tag{2}$$

where $\mathbf{a}_n = [A_n \cdots A_{n-p+1}]^T$ is the p-dimensional state vector and $\mathbf{v}_n$ denotes the zero-mean prediction residual with covariance matrix $\mathbf{Q}_n$. The $(p \times p)$ transition matrix has the form $\mathbf{F}_n = \begin{bmatrix} -\mathbf{b}_n^T \\ \mathbf{I} \quad \mathbf{0} \end{bmatrix}$, where $\mathbf{b}_n = [b_1 \cdots b_p]^T$ is the vector of linear prediction (LPC) coefficients for the speech spectral amplitudes in frame n. Our model differs from that used in [12] in two respects: we treat the noise and speech as additive in the complex STFT domain rather than in the spectral amplitude domain and we use a generalized Gamma prior for the speech amplitudes rather than a Gaussian prior.

3. PROPOSED ESTIMATOR DESCRIPTION

A block diagram of the proposed algorithm is shown in Fig. 1. The noisy speech, $y(t)$ is converted to the time-frequency domain, $R_{n,k}e^{j\Theta_{n,k}}$, using the STFT [15]. In order to perform LPC modelling in the modulation domain, the noise power spectrum is estimated using, for example, [16] or [17], and the speech is passed through a conventional MMSE enhancer [3] to reduce the effects of the noise on the modelling. Following this, the sequence of spectral amplitudes in each frequency bin is divided into overlapping modulation frames. Autocorrelation LPC [18] is performed on each modulation frame to determine the coefficients, $\mathbf{b}_n$, and thence the transition matrix $\mathbf{F}_n$ in (2).

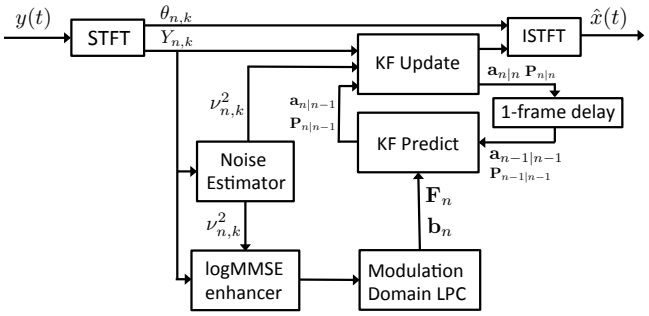

Fig. 1. Block diagram of the proposed Kalman Filter MMSE estimator.

3.1. Kalman filter prediction step

From the time update model (2), we obtain the KF prediction equations

$$\mathbf{a}_{n|n-1} = \mathbf{F}_{n-1}\mathbf{a}_{n-1|n-1} \tag{3}$$

$$\mathbf{P}_{n|n-1} = \mathbf{F}_{n-1}\mathbf{P}_{n-1|n-1}\mathbf{F}^T_{n-1} + \mathbf{Q}_n, \tag{4}$$

where $\mathbf{a}_{n|n-1}$ and $\mathbf{P}_{n|n-1}$ denote respectively the a priori estimates of the amplitude state vector and of the corresponding covariance matrix at time n, and $\mathbf{a}_{n-1|n-1}$ denotes the a posteriori estimate of the state vector at time $n-1$. The first element of the state vector, $\mathbf{a}_{n|n-1}$, corresponds to the spectral amplitude in the current frame, $A_{n|n-1}$, and so its a priori mean and variance are given by

$$\mu_{n|n-1} \triangleq E(A_n|\mathcal{R}_{n-1}) = \mathbf{d}^T\mathbf{a}_{n|n-1} \tag{5}$$

$$\sigma^2_{n|n-1} \triangleq Var(A_n|\mathcal{R}_{n-1}) = \mathbf{c}^T\mathbf{P}_{n|n-1}\mathbf{c}, \tag{6}$$

where $\mathcal{R}_{n-1}$ represents the observed speech amplitudes up to time $n-1$ and $\mathbf{c} = [1\,0...0]^T$.

3.2. Kalman Filter MMSE update model

In this section, we describe the KF MMSE update step which determines an updated state estimate by combining the predicted state vector and covariance, the estimated noise and the observed spectral amplitude. Within the update step, we model the prior speech amplitude $A_{n|n-1}$ using a 2-parameter Gamma distribution

$$p\left(a_n|\mathcal{R}_{n-1}\right) = \frac{2a_n^{2\gamma_n-1}}{\beta_n^{2\gamma_n}\Gamma\left(\gamma_n\right)}\exp\left(-\frac{a_n^2}{\beta_n^2}\right), \tag{7}$$

where $\Gamma\left(\cdot\right)$ is the Gamma function. The distribution is obtained by setting $c = 2$ in the generalized Gamma distribution given in [19], and the two parameters, β_n and γ_n are chosen to match the mean μ_n and variance σ_n^2 of the predicted amplitude from (5) and (6). Examples of the probability density functions from (7) with variance, $\sigma^2 = 1$ and means, μ, in the range 0.5 to 8 are shown in Fig. 2, from which it can be seen that the distribution in (7) is sufficiently flexible to model the outcome of the prediction over a wide range of μ_n/σ_n.

At frame n, the mean and variance of the Gamma distribution in (7) can be expressed in terms of β_n and γ_n [19] as

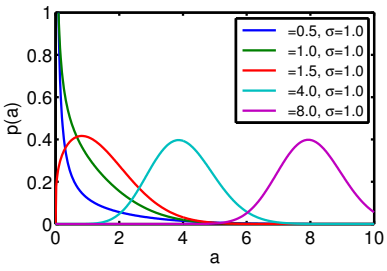

Fig. 2. Curves of Gamma probability density function for (7) with variance $\sigma^2 = 1$ and different means.

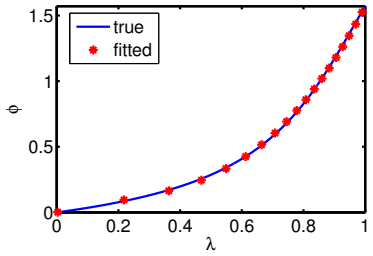

Fig. 3. The curve of ϕ versus λ, where $0 < \phi = \arctan(\gamma) < \frac{\pi}{2}$ and $0 < \lambda = \frac{\Gamma^2(\gamma+0.5)}{\Gamma^2(\gamma)\gamma} < 1$.

$$\mu_{n|n-1} = \beta_n\frac{\Gamma\left(\gamma_n+0.5\right)}{\Gamma\left(\gamma_n\right)}, \tag{8}$$

$$\sigma^2_{n|n-1} = \beta_n^2\left(\gamma_n - \frac{\Gamma^2\left(\gamma_n+0.5\right)}{\Gamma^2\left(\gamma_n\right)}\right). \tag{9}$$

We can eliminate β between (8) and (9) to obtain

$$\frac{\Gamma^2\left(\gamma_n+0.5\right)}{\gamma_n\Gamma^2\left(\gamma_n\right)} = \frac{\mu^2_{n|n-1}}{\mu^2_{n|n-1}+\sigma^2_{n|n-1}} \triangleq \lambda_n \tag{10}$$

We need to solve the non-linear equation (10) to determine γ_n from the value of λ_n which can be calculated from $\mu_{n|n-1}$ and $\sigma^2_{n|n-1}$ and satisfies $0 < \lambda_n < 1$. Instead of dealing with γ_n directly, it is convenient to set $\phi_n = \arctan(\gamma_n)$ where ϕ_n lies in the range $0 < \phi_n < \frac{\pi}{2}$. The solid line in Fig. 3 shows the function $\phi_n(\lambda)$. We can approximate this function well with a low-order polynomial that is constrained to pass through the points $(0,0)$ and $(1, \frac{\pi}{2})$ and in the experiments in Sec. 4.1 we use the quartic approximation

$$\phi(\lambda) = -0.1640\lambda^4 + 2.3612\lambda^3 - 1.2182\lambda^2 + 0.5918\lambda$$

which is shown with asterisks in Fig. 3. Given λ_n we can use this polynomial to obtain ϕ_n and thence γ_n by the inverse transform $\gamma_n = \tan(\phi_n)$.

3.3. Derivation of estimator

The MMSE estimate of A_n is given by the conditional expectation

$$\mu_{n|n} = E(A_n|\mathcal{R}_n) = \int_0^\infty a_n p(a_n|\mathcal{R}_n)da_n \tag{11}$$

Using Bayes rule, the conditional probability is expressed as

$$p\left(a_n|\mathcal{R}_n\right) = p(a_n|y_n,\mathcal{R}_{n-1}) = \frac{\int_0^{2\pi} p\left(y_n|a_n,\phi_n,\mathcal{R}_{n-1}\right) p\left(a_n,\phi_n|\mathcal{R}_{n-1}\right) d\phi_n}{p\left(y_n|\mathcal{R}_{n-1}\right)} \tag{12}$$

where ϕ_n is the realization of the random variable Φ_n which represents the phase of the clean speech. Because Y_n is conditionally independent of $\mathcal{R}_{n-1}$ given a_n and ϕ_n, (12) becomes

$$p\left(a_n|\mathcal{R}_n\right) = \frac{\int_0^{2\pi} p\left(y_n|a_n,\phi_n\right) p\left(a_n,\phi_n|\mathcal{R}_{n-1}\right) d\phi_n}{p\left(y_n|\mathcal{R}_{n-1}\right)} \tag{13}$$

Following [2], the observation noise is assumed to be complex Gaussian distributed with variance $\nu_n^2 = E(N_n^2)$ leading to the observation prior model

$$p(y_n|a_n,\phi_n) = \frac{1}{\pi\nu_n^2}\exp\left\{-\frac{1}{\nu_n^2}|y_n - a_n e^{j\phi_n}|^2\right\} \tag{14}$$

Under the assumption of the statistical models previously defined and assuming that the phase components and amplitude components, Φ_n and A_n, are independent, we can now calculate a closed-form expression for the estimator (11) using [20, Eq. 6.631, 9.201.1, 9.220.2]

$$\mu_{n|n} = \frac{\Gamma\left(\gamma_n + 0.5\right)}{\Gamma\left(\gamma_n\right)}\sqrt{\frac{\xi_n}{\zeta_n(\gamma_n+\xi_n)}}\frac{\mathcal{M}\left(\gamma_n+0.5;1;\frac{\zeta_n\xi_n}{\gamma_n+\xi_n}\right)}{\mathcal{M}\left(\gamma_n;1;\frac{\zeta_n\xi_n}{\gamma_n+\xi_n}\right)}R_n, \tag{15}$$

where $\mathcal{M}$ is the confluent hypergeometric function [21], and

$$\xi_n = \frac{E(A_n^2|\mathcal{Y}_{n-1})}{\nu_n^2} = \frac{\mu_{n|n-1}^2 + \sigma_{n|n-1}^2}{\nu_n^2},\ \zeta_n = \frac{R_n^2}{\nu_n^2}$$

are the a priori SNR and a posteriori SNR respectively. The variance of the posterior estimate is given by

$$\sigma_{n|n}^2 = \mathrm{E}\left(A_n^2|\mathcal{R}_n,\phi_n\right) - \left(\mathrm{E}\left(A_n|\mathcal{R}_n,\phi_n\right)\right)^2 = \frac{\gamma_n\xi_n}{\zeta_n(\gamma_n+\xi_n)}\frac{\mathcal{M}\left(\gamma_n+1;1;\frac{\zeta_n\xi_n}{\gamma_n+\xi_n}\right)}{\mathcal{M}\left(\gamma_n;1;\frac{\zeta_n\xi_n}{\gamma_n+\xi_n}\right)}R_n^2 - \left(\mu_{n|n}\right)^2. \tag{16}$$

3.4. Update of state vector

The final step is to update the entire state vector and the associated covariance matrix, $\mathbf{a}_{n|n}$ and $\mathbf{P}_{n|n}$. In order to decorrelate the current observation from the rest of the state vector, we decompose the covariance matrix $\mathbf{P}_{n|n-1}$ as

$$\mathbf{P}_{n|n-1} = \begin{bmatrix} \sigma_{n|n-1}^2 & \mathbf{g}_n^T \\ \mathbf{g}_n & \mathbf{G}_n \end{bmatrix},$$

where $\mathbf{g}_n$ is a $(p-1)$-dimensional vector. We now transform the state vector as

$$\mathbf{z}_{n|n-1} = \mathbf{H}_n\mathbf{a}_{n|n-1} \tag{17}$$

using the transformation matrix $\mathbf{H}_n = \begin{bmatrix} 1 & \mathbf{0}^T \\ -\frac{\mathbf{g}_n}{\sigma_{n|n-1}^2} & \mathbf{I} \end{bmatrix}$. The covariance matrix, $\mathbf{U}_{n|n-1}$, of the transformed state vector $\mathbf{z}_{n|n-1}$ is given by

$$\mathbf{U}_{n|n-1} = E\left(\mathbf{z}_{n|n-1}\mathbf{z}_{n|n-1}^T\right) = \mathbf{H}_n\mathbf{P}_{n|n-1}\mathbf{H}_n^T = \begin{bmatrix} \sigma_{n|n-1}^2 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{G}_n - \sigma_{n|n-1}^{-2}\mathbf{g}_n\mathbf{g}_n^T \end{bmatrix}.$$

We see that the first element of $\mathbf{z}_{n|n-1}$ is equal to $\mu_{n|n-1}$ and uncorrelated with any of the other elements and is therefore distributed as $\mathcal{N}(\mu_{n|n-1},\sigma_{n|n-1}^2)$. Using the posterior mean and variance from (15) and (16) and $\mathbf{c} = [1\,0\ldots0]^T$, we can update the transformed mean vector and covariance matrix as

$$\mathbf{z}_{n|n} = \mathbf{z}_{n|n-1} + (\mu_{n|n} - \mu_{n|n-1})\mathbf{c}$$
$$\mathbf{U}_{n|n} = \mathbf{U}_{n|n-1} + \left(\sigma_{n|n}^2 - \sigma_{n|n-1}^2\right)\mathbf{c}\mathbf{c}^T.$$

Inverting the transformation in (17), we obtain, after some algebraic manipulation, the following update equations

$$\mathbf{a}_{n|n} = \mathbf{a}_{n|n-1} + \left(\mu_{n|n} - \mu_{n|n-1}\right)\sigma_{n|n-1}^{-2}\mathbf{P}_{n|n-1}\mathbf{c} \tag{18}$$

$$\mathbf{P}_{n|n} = \mathbf{P}_{n|n-1} + \left(\sigma_{n|n}^2\sigma_{n|n-1}^{-2} - 1\right)\sigma_{n|n-1}^{-2}\mathbf{P}_{n|n-1}\mathbf{c}\mathbf{c}^T\mathbf{P}_{n|n-1}. \tag{19}$$

In this section we have derived the update equations for the KF. For each acoustic frame of noisy speech, we first use (3) and (4) to calculate the a priori state vector $\mathbf{a}_{n|n-1}$ and the corresponding covariance $\mathbf{P}_{n|n-1}$, and solve (10) to find γ_n. We then use (15) and (16) to calculate the a posteriori estimate of the amplitude and the corresponding variance respectively. Finally, the KF state vector and its covariance matrix are updated using (18) and (19).

4. IMPLEMENTATION AND EVALUATION

4.1. Implementation of algorithm

In this section, we compare the performance of the proposed KF based MMSE (KMMSE) estimator with five other algorithms: (i) logMMSE – the baseline log-amplitude MMSE enhancer from [3, 22]; (ii) pMMSE – the perceptually motivated MMSE estimator from [23, 22] using a weighted Euclidean distortion measure with a power exponent of $p = -1$; (iii) ModSub – the modulation-domain spectral subtraction from [11]; (iv) MDKF – the version of the modulation-domain Kalman filter from [12] that extracts the modulation-domain LPC coefficients from enhanced speech (using the logMMSE algorithm [3, 22]); (v) KFMMSEI – an intermediate version of our proposed algorithm that assumes the speech and noise add in the STFT amplitude domain rather than the complex STFT domain (i.e. replacing (1) with $|Y_{n,k}| = |X_{n,k}| + |W_{n,k}|$). The parameters of all the algorithms were chosen to optimize performance on a subset of the training set of the TIMIT database [24]. We have used an acoustic frame length of 32 ms with a 4 ms increment which gives a 250 Hz sampling frequency in the modulation domain. The speech LPC models are determined from a modulation frame of duration 128 ms (32 acoustic frames) with a 16 ms frame increment and the model orders in both the KMMSE and MDKF algorithms are $l = 2$. In the experiments, we use the core test set from the TIMIT database which contains 16 male and 8 female speakers

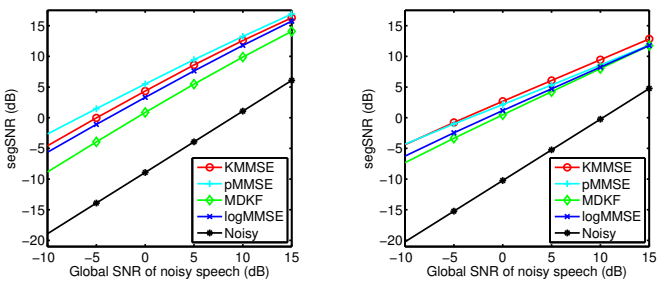

Fig. 4. Average segmental SNR of enhanced speech speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive car noise (left) and street noise (right). The algorithm acronyms are defined in the text.

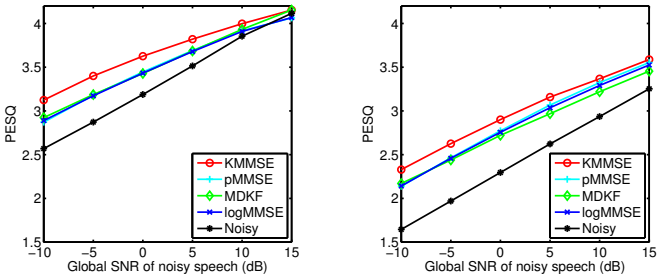

Fig. 5. Average PESQ quality of enhanced speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive car noise (left) and street nose (right).

each reading 8 distinct sentences (totalling 192 sentences) and the speech is corrupted by the noise from the RSG-10 database [25] and the ITU-T test signals database [26] at $-10, -5, 0, 5, 10$ and 15 dB global SNR. A Hamming window is used in the STFT analysis and synthesis and the noise power spectrum, $\nu_{n,k}^2$, is estimated using the algorithm from [17] as implemented in [22]. It is possible for the algorithm to lock up with $\mu_{n|n} = 0$; to prevent this, we impose the constraint $\gamma_n > 0.5$ in (10).

4.2. Performance evaluations

The performance of the algorithms is evaluated using both segmental SNR (segSNR) and the perceptual evaluation of speech quality (PESQ) measure defined in ITU-T P.862. All the measured values shown are averages over the 192 sentences in the TIMIT core test set. Figure 4 shows the average segSNR of speech enghanced by the proposed algorithm (KMMSE) as well as by the logMMSE, pMMSE and MDKF algorithms. The left and right plots respectively show results for car noise [25] and street noise [26]. We see that for car noise, which is predominantly low frequency, pMMSE gives the best segSNR especially at poor SNRs where it is approximately 2 dB better than KMMSE, the next best algorithm. For street noise however, which has a broader spectrum, the situation is reversed and the KMMSE algorithm has the best performance especially at SNRs above 5 dB. Figure 5 shows the corresponding average PESQ scores for car noise (left plot) and street noise (right plot). We see that, with this measure, the KMMSE algorithm clearly has the highest performance. For car noise, the PESQ score from the KMMSE algorithm is approximately 0.2 better than that of the other algorithms at SNRs below 5 dB while for street noise, the corresponding figure is 0.15. These differences correspond to SNR improvements of 4 dB and 2.5 dB respectively. To assess the robustness to noise type, we have evaluated the algorithms using twelve different noise types from [25] with the average SNR for each noise type chosen to give a mean PESQ score of 2.0 for the noisy speech. In Fig. 6, the solid lines show the median, the boxes the interquartile range and the whiskers the extreme PESQ values for the 198×12 speech-plus-noise combinations. Figure 7 shows box plots of the difference in PESQ score between competing algorithms and KMMSE. We see that in all cases the entire box lies below the axis line; this indicates that KMMSE results in an improvement for an overwhelming majority of speech-plus-noise combinations. The KMMSEI box plot demonstrates the small but consistent benefit of using an additive model in the complex STFT domain rather than the amplitude domain.

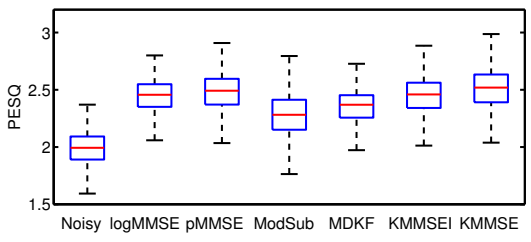

Fig. 6. Box plot of the PESQ scores for noisy speech processed by six enhancement algorithms. The plots show the median, interquartile range and extreme values from 2376 speech+noise combinations.

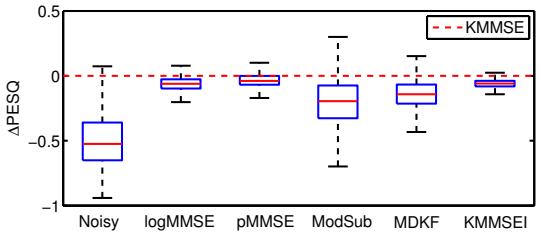

Fig. 7. Box plot showing the difference in PESQ score between competing algorithms and the proposed algorithm, KMMSE for 2376 speech+noise combinations.

5. CONCLUSION

In this paper we have proposed an MMSE spectral amplitude estimator based on a modulation domain Kalman filter. The novel MMSE estimator incorporates a model of the temporal dynamics of spectral amplitudes within each frequency bin by using a Kalman filter. We have shown how the parameters of the speech prior model can be estimated from the predicted state vector from the Kalman filter, and used to calculate the estimator in the update step. The proposed algorithm gives a consistent improvement in PESQ over all the competitive algorithms demonstrating that PESQ can be improved by about 0.2 over the baseline logMMSE enhancer for a wide range of SNRs.

6. REFERENCES

[1] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Process.*, 27(2):113 – 120, April 1979.

[2] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(6):1109–1121, December 1984.

[3] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 33(2):443–445, 1985.

[4] L. Atlas and S.A. Shamma. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, 7:668–675, 2003.

[5] R. Drullman, J. M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2):1053–1064, 1994.

[6] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.

[7] B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech communication*, 25(1):117–132, 1998.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time frequency weighted noisy speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(7):2125–2136, September 2011.

[9] R. L. Goldsworthy and J. E. Greenberg. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *J. Acoust. Soc. Am.*, 116(6):3679–3689, December 2004.

[10] T. H. Falk, S. Stadler, W. B. Kleijn, and W. Y. Chan. Noise suppression based on extending a speech-dominated modulation band. In *Proc. Interspeech Conf.*, pages 970–973, August 2007.

[11] K. Paliwal, K. Wojcicki, and B. Schwerin. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Communication*, 52(5):450–475, 2010. The Matlab software is available online at URL: http://maxwell.me.gu.edu.au/spl/research/modspecsub/.

[12] S. So and K. K. Paliwal. Modulation-domain Kalman filtering for single-channel speech enhancement. *Speech Communication*, 53(6):818–829, July 2011.

[13] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Trans. Speech Audio Process.*, 15(6):1741–1752, 2007.

[14] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Process.*, 13(5):845–856, September 2005.

[15] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete Time Processing of Speech Signals*. Prentice Hall, 1993.

[16] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.*, 9(5):504 –512, July 2001.

[17] T. Gerkmann and R. C. Hendriks. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(4):1383–1393, May 2012.

[18] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561 – 580, April 1975.

[19] L. Norman, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. Wiley, 1994.

[20] A. Jeffrey and D. Zwillinger. *Table of Integrals, Series, and Products*. Academic Press, 2007.

[21] F. Olver, D. Lozier, R. F. Boiszert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions: Companion to the Digital Library of Mathematical Functions*. Cambridge University Press, 2010. URL: http://dlmf.nist.gov/13.

[22] D. M. Brookes. VOICEBOX: A speech processing toolbox for MATLAB. http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1998-2014.

[23] P. C. Loizou. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. *IEEE Trans. Speech Audio Process.*, 13(5):857–869, 2005.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. TIMIT acoustic-phonetic continuous speech corpus. Corpus LDC93S1, Linguistic Data Consortium, Philadelphia, 1993.

[25] H. J. M. Steeneken and F. W. M. Geurtsen. Description of the RSG.10 noise data-base. Technical Report IZF 1988–3, TNO Institute for perception, 1988.

[26] ITU-T P.501. Test signals for use in telephonometry, August 1996.