MODULATION-DOMAIN SPEECH ENHANCEMENT USING A KALMAN FILTER WITH A BAYESIAN UPDATE OF SPEECH AND NOISE IN THE LOG-SPECTRAL DOMAIN

Nikolaos Dionelis and Mike Brookes

Dept. of Electrical and Electronic Engineering, Imperial College London, UK

ABSTRACT

We present a Bayesian estimator that performs log-spectrum estimation of both speech and noise, and is used as a Bayesian Kalman filter update step for single-channel speech enhancement in the modulation domain. We use Kalman filtering in the log-power spectral domain rather than in the amplitude or power spectral domains. In the Bayesian Kalman filter update step, we define the posterior distribution of the clean speech and noise log-power spectra as a two-dimensional multivariate Gaussian distribution. We utilize a Kalman filter observation constraint surface in the three-dimensional space, where the third dimension is the phase factor. We evaluate the results of the phase-sensitive log-spectrum Kalman filter by comparing them with the results obtained by traditional noise suppression techniques and by an alternative Kalman filtering technique that assumes additivity of speech and noise in the power spectral domain.

Index Terms— Speech enhancement; noise suppression

1. INTRODUCTION

Single-channel speech enhancement in non-stationary noise environments remains a challenging task since algorithms encounter the tradeoff between noise reduction and speech distortion. An overview of statistical-based algorithms is given in [1]. Whereas model-based statistical algorithms treat each time-frame independently, an alternative approach performs filtering in the modulation domain, which models the temporal/inter-frame correlation of speech and utilizes information in speech that is carried by the modulation of the spectral envelopes rather than by the envelopes themselves [2] [3].

Modulation-domain Kalman filtering refers to sequentially updating the statistics of clean speech using a Kalman filter (KF) prediction step that involves modulation frames. The algorithms in [3] [4] operate in the modulation domain, use overlapping modulation frames and the KF. In [5] [6], KF tracking of speech (and noise in [7]) is presented. Based on [2] and [8], researchers perform speech Kalman filtering in the amplitude spectral domain assuming additivity of speech and noise in the amplitude domain, which is a deliberate approximation, and assuming that speech and noise follow Gaussian distributions in the amplitude domain. In [2], speech-noise additivity in the amplitude domain assumes that the phase factor, the cosine of the phase asynchrony between speech and noise, is unity.

Considering KF-related algorithms, many papers, such as [9], [10] and [11], use the observation model and the non-linear log-spectral distortion equation relating clean speech and noisy speech in the log-power spectral domain or in other spectral domains. Specifically, in [9] and [10], the non-linear environment distortion model in the cepstral domain is utilized. Amongst other technical papers, the non-linear log-power distortion equation is used in [12] and in [13].

The algorithms in [9] and [10], which are also described in [12], re-estimate the distortion parameters of the noise mean and variance using a variant of the EM algorithm. Based on [10], the noise re-estimation heuristic is not successful in low SNR levels and in non-stationary noise conditions. The iterative re-estimation heuristic of the noise mean and variance in [10] would not be needed if the noise posterior distribution given the noisy observation was defined.

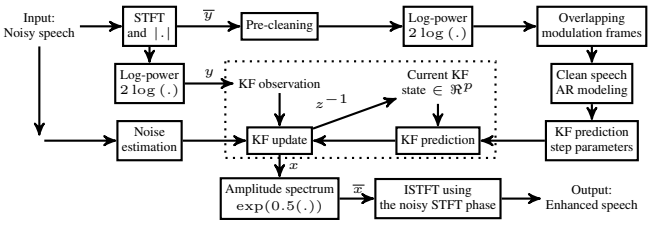

Fig. 1. The flowchart of the algorithm. The term z^{-1} refers to one-frame delay. The blocks in the dotted rectangle constitute the speech Kalman filter (KF) and, in Fig. 2, the KF over time t is illustrated.

A phase-sensitive model is employed in the non-linear log-spectral distortion equation in [14]. The phase factor is the cosine of the difference between the clean speech and noise STFT phases. At local SNR levels around 0 dB, the phase term should not be neglected [13] [15]. Amongst other technical papers, the phase factor is used in [12] and in Sec 4.8.2 of [14]. Many algorithms, such as logNMF [16] due to the max approximation method, ignore the phase factor and assume the additivity of speech and noise powers.

In this paper, as main innovation, we create a speech-noise KF in the log-power spectral domain and we advance the KF update by performing Bayesian estimation of both speech and noise. We approximate the non-Gaussian posterior distribution of the clean speech and noise log-power spectra with a Gaussian distribution, using the theoretical distribution of the phase factor. In the two-dimensional Gaussian, the off-diagonal term of the state covariance matrix models the correlation between speech and noise log-powers. Using the KF state as the log-power spectrum of clean speech, we minimize the log-power spectrum estimation error. Having a Gaussian in the non-nonnegative log-power domain leads to good speech modeling since researchers use super-Gaussian distributions that resemble the log-normal, such as the Gamma [3] [4], in the amplitude domain.

2. THE SPEECH ENHANCEMENT ALGORITHM

The flowchart of the algorithm is shown in Fig. 1. The algorithm's first step is to perform windowing and the short time Fourier transform (STFT). After the STFT, we perform three different actions: we first do pre-cleaning and estimate an autoregressive (AR) model for clean speech using modulation frames. We secondly do (observation) noise power spectrum estimation using an existing algorithm;

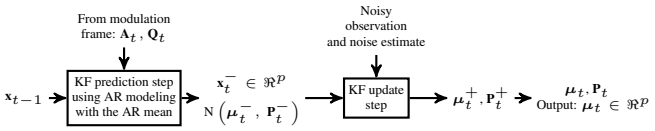

Fig. 2. The KF is shown. The inputs and outputs in this diagram match the inputs and outputs of the dotted rectangle in Fig. 1. We see how the speech Gaussian KF state evolves/changes at time t.

Fig. 1 shows the algorithm without KF noise tracking. Thirdly, we perform speech Kalman filtering in the log-power spectral domain.

In Fig. 1, the blocks in the dotted rectangle constitute the KF. The inputs to the KF are: the noisy speech in the log-power STFT spectral domain that constitutes the KF observation, the estimated noise log-power and the KF transition matrix $\mathbf{A}_t$ that is created from autoregressive AR modeling of order p on the pre-cleaned modulation frame. The output of the dotted rectangle in Fig. 1 is the p-dimensional KF state. The algorithm's next step is to keep the first element of the KF state mean that is the estimated speech log-power spectrum, and transform it to the amplitude spectral domain. Finally, we reconstuct the estimated clean speech signal in the time domain using the inverse STFT (ISTFT) with the noisy STFT phase.

2.1. The signal model and the speech KF prediction step

We assume that in the complex STFT domain, the noisy speech is:

$$\bar{y}_t(k)e^{j\theta_t(k)} = \bar{x}_t(k)e^{j\phi_t(k)} + \bar{n}_t(k)e^{j\psi_t(k)} \tag{1}$$

In (1), we use the time index t and the frequency index k and, from now on, we omit k. The amplitudes of the noisy speech, clean speech and noise are respectively $\bar{y}$, $\bar{x}$ and $\bar{n}$. The noisy speech phase is θ, the clean speech phase is ϕ and the noise phase is ψ. The log-powers of the noisy speech, clean speech and noise are respectively $y = 2\log(\bar{y})$, $x = 2\log(\bar{x})$ and $n = 2\log(\bar{n})$. In Fig. 1, we denote the noisy and clean speech log-powers and amplitudes.

Figure 2 shows the speech KF state at time step t before and after the KF prediction and update steps. We use the linear KF prediction equations in (2) based on AR modeling [17]. In (2), the transition noise $\mathbf{w}_{\mathbf{Q}_t}$ is zero-mean Gaussian, the KF transition noise covariance matrix is $\mathbf{Q}_t$ and the KF transition matrix is $\mathbf{A}_t$. The vector of speech AR coefficients is $\mathbf{a}_t$ and the AR modeling error variance is q.

$$\mathbf{x}_t^- = \mathbf{A}_t\mathbf{x}_{t-1} + \mathbf{w}_{\mathbf{Q}_t}, \quad \mathbf{x}_t = \left(x_t\ x_{t-1}\ \ldots\ x_{t-p+1}\right)^T \in \Re^p \tag{2}$$
$$\mathbf{A}_t = \begin{pmatrix} -\mathbf{a}_t^T \\ \mathbf{I}\ \mathbf{0} \end{pmatrix}, \mathbf{Q}_t = \begin{pmatrix} q & \mathbf{0} \\ 0 & \mathbf{0} \end{pmatrix}, \quad \mathbf{A}_t, \mathbf{Q}_t \in \Re^{p\times p}$$

As seen in Figs. 1-2 and (2), $\mathbf{A}_t$ and $\mathbf{Q}_t$ are created from AR(p) modeling on the pre-cleaned modulation frame using the covariance method [18] [19], estimating the AR coefficients and the AR mean.

2.2. KF noise tracking and the joint speech-noise KF state

We now use a joint speech-noise state with a full covariance matrix. As in (2) and in Sec. 2.1, we do KF noise tracking in the log-power spectral domain based on AR(r) modeling and on the estimated SNR in the modulation frame [3]. After the noise KF prediction step, we decorrelate the joint KF state and, then, we multiply the noise log-power Gaussian with the Gaussian that is obtained from external noise estimation and log-normal noise power modeling [20] [21].

As in (2), we now use a joint speech-noise state $\mathbf{x}_t^{(j)} \in \Re^{p+r}$ in (3). The superscripts (j) and (n) denote the joint speech-noise KF state/parameters and the noise KF state/parameters respectively.

$$\mathbf{x}_t^{(j)-} = \mathbf{A}_t^{(j)}\mathbf{x}_{t-1}^{(j)} + \mathbf{w}_{\mathbf{Q}_t}^{(j)}, \quad \mathbf{A}_t^{(n)}, \mathbf{Q}_t^{(n)} \in \Re^{r\times r},\ \mathbf{x}_t^{(n)} \in \Re^r \tag{3}$$
$$\mathbf{A}_t^{(j)} = \begin{pmatrix} \mathbf{A}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_t^{(n)} \end{pmatrix},\ \mathbf{Q}_t^{(j)} = \begin{pmatrix} \mathbf{Q}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_t^{(n)} \end{pmatrix} \in \Re^{(p+r)\times(p+r)}$$

2.3. The phase factor and the Bayesian KF update step

The Bayesian KF update step estimates the posterior distribution of clean speech and noise log-powers given the noisy observation log-power. The KF update step considers the Gaussian clean speech and noise prior distributions from the KF prediction step, the distribution of the STFT phase difference between clean speech and noise and the observation constraint surface in the three-dimensional space, where the third dimension is the phase factor $\alpha \in [-1, 1]$ [13]. The phase factor $\alpha = \cos(\phi - \psi)$ affects the power spectral domain [13]:

$$\bar{y}^2 = \bar{x}^2 + \bar{n}^2 + 2\alpha\,\bar{n}\bar{x} \tag{4}$$

Based on Sec. 5.2 in [22], for a uniform phase difference, the phase factor α distribution is: $p(\alpha) = \left(\pi\sqrt{1-\alpha^2}\right)^{-1}$ for $-1 < \alpha < 1$, and $E\{\alpha^n\} = \dfrac{2^{-n}\times n!}{\left((0.5n)!\right)^2}$ for even n and zero otherwise.

We work in the complex STFT domain. We have a prior distribution for $p(x, n, \alpha)$ and we wish to apply a constraint on y to get a posterior distribution. Using log-powers, as in Sec. 2.1, we obtain:

$$e^y = e^x + e^n + 2e^{0.5(x+n)}\alpha \tag{5}$$

From which: $\alpha = \dfrac{e^y - e^x - e^n}{2\times e^{0.5(x+n)}}$. We use: $u = n - x$, $v = n + x$, $\alpha = 0.5\exp(y - 0.5v) - \cosh(0.5u)$. Using u, v, α and $v = 2\left(y - \log\left(2(\alpha + \cosh(0.5u))\right)\right)$, we change variables $v \Rightarrow y$:

$$\begin{pmatrix} u \\ y \\ \alpha \end{pmatrix} = \begin{pmatrix} u \\ 0.5v + \log\left(2(\alpha + \cosh(0.5u))\right) \\ \alpha \end{pmatrix}$$

We do the variable transformation from (u, v, α) to (u, y, α). The Jacobian determinant is $\Delta = 0.5$. From this, the posterior is:

$$p(u, \alpha\,|\,y) = \left.\frac{p(u, \alpha,\ y)}{p(y)}\right|_y \propto \left(p(u, v)p(\alpha) \times |\Delta|^{-1}\right)\big|_y \propto p(\alpha)\times$$
$$\times\ \mathscr{N}\left(\begin{pmatrix} u \\ 2\left(y - \log\left(2(\alpha + \cosh(0.5u))\right)\right) \end{pmatrix};\ \mathbf{m}_u,\ \mathbf{S}_u\right)$$

In the preceding equation, we use a two-dimensional Gaussian distribution for $p(u, v)$ with mean $\mathbf{m}_u$ and covariance matrix $\mathbf{S}_u$.

Next, we use $E\{u^a v^b | y\}$ to calculate the first and second moments of the posterior distribution of (u, v), which determines (x, n). The aim is to compute $E\{x^a n^b | y\}$ where $0 \le a + b \le 2$.

$$E\{u^a v^b\,|\,y\} = \int_{\alpha=-1}^{1}\int_{u=-\infty}^{\infty} u^a v^b\, p(u, \alpha\,|\,y)\, du\, d\alpha =$$
$$= \frac{1}{|\Delta|\, p(y)} \int_{\alpha=-1}^{1} p(\alpha) \int_{u=-\infty}^{\infty} u^a v^b\, p(u, v)\, du\, d\alpha \tag{6}$$

In (6), the inner integration over u is performed with straight line segments and truncated Gaussians, thus obtaining a closed-form solution. In (6), the outer integration over α is done using n weighted sigma points [23] [24]. We use the Unscented transform [25] [26] and $n = 3$ weighted sigma points for approximating integration with summation. We fit the first three even moments of the phase factor α using (7). The outer integral in (6) is exact to $f(\alpha)$ polynomials up to the fifth order in (7) when we use three sigma points.

$$E\{f(\alpha)\} = \int_{\alpha=-\infty}^{\infty} p(\alpha)\, f(\alpha)\, d\alpha = \sum_{i=1}^{n} w_i\, f(\alpha_i) \qquad (7)$$

We now use the left-hand and right-hand side expressions of the equation in (7). We utilize $n = 3$ sigma points and we have:

$$1 = \sum_{i=1}^{3} w_i, \quad 0.5 = \sum_{i=1}^{3} w_i\, a_i^2, \quad 0.3750 = \sum_{i=1}^{3} w_i\, a_i^4$$

$$w_1 = w_2 = w_3 = \frac{1}{3},\ a_1 = \frac{\sqrt{3}}{2},\ a_2 = -\frac{\sqrt{3}}{2},\ a_3 = 0$$

For n sigma points, we observe that $w_i = n^{-1}, i \in \{1, 2, \ldots, n\}$. We now integrate over α in (6) by summing over the $n = 3$ sigma points using (7). In (7), noting that $v = v(u, \alpha)$, we compute:

$$f(\alpha) = \int_{u=-\infty}^{\infty} u^a v^b \mathcal{N}\left(\begin{pmatrix} u \\ v \end{pmatrix}; \mathbf{m}_u, \mathbf{S}_u\right) du \qquad (8)$$

In the KF update, we update the mean and covariance matrix of the current speech-noise KF state. We calculate the first and second moments of the posterior distribution of the current speech and noise log-power spectra using $E\{x^a n^b | y\}$ where $0 \leq a + b \leq 2$. We utilize $p(u, v)$ instead of $p(x, n)$ since (u, v) uniquely defines (x, n), which is the current speech-noise KF state. In (6), we find the moments to find the posterior of the speech-noise KF state in the rotated u-v domain. In (6)-(8), we use the six cases with $0 \leq a + b \leq 2$.

In the phase-sensitive KF update step, we compute the two-dimensional Gaussian posterior distribution of clean speech and noise log-power spectra using sigma points for integration over α, considering the phase difference between speech and noise. We use the phase factor α with the KF observation employing an observation constraint surface for $\alpha \in [-1, 1]$, rather than a constraint line for $\alpha = 0$ (or $\alpha = 1$), in the three-dimensional space of x, n, α.

Figure 3 shows the posterior distribution when the prior distribution is in the curvy triangle, which is defined in (9), and on the $u = 0$ line that corresponds to SNR $= 0$ dB. Likewise, Fig. 4 illustrates the posterior distribution when the prior distribution is out of the curvy triangle and on the $u = 0$ line and Fig. 5 when the prior distribution is in the curvy triangle and off the $u = 0$ line. Figure 6 depicts two cases when the prior distribution is out of the curvy triangle and off the $u = 0$ line. In Fig. 6, prior 1 is at the upper left side out of the curvy triangle and thus at a negative SNR position. The background in Figs. 3-6 is based on the Gaussian posterior covariance matrix.

Figure 5 is the most important from Figs. 3-6 since it examines a frequent positive SNR case. Based on Fig. 5, the algorithm works as expected. In prior 2 of Fig. 6, we have a rare positive SNR case. The equations of the curvy triangle in the u-v domain are (9). The surface of the curvy triangle is related to the different values of α.

$$\begin{aligned} \alpha = 1, &\quad \cosh(0.25u) = 0.5 \exp(-0.25v) \\ \alpha = -1, &\quad |\sinh(0.25u)| = 0.5 \exp(-0.25v) \end{aligned} \qquad (9)$$

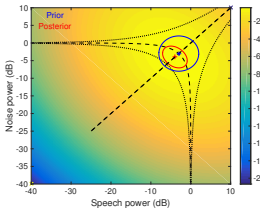

Fig. 3. Plot of the KF posterior distribution, the prior distribution, the $a = -1, 0, 1$ constraints and the $u = 0$ line. The prior distribution is in the curvy triangle and on the $u = 0$ line.

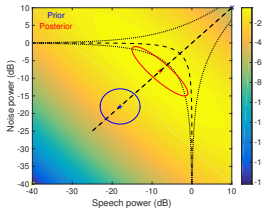

Fig. 4. Plot of the KF posterior distribution and the prior distribution. In this case, the prior distribution is out of the curvy triangle and on the $u = 0$ line, which indicates an SNR of 0 dB.

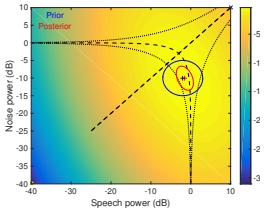

Fig. 5. Plot of the KF posterior distribution and the prior distribution. In this case, the prior distribution is in the curvy triangle and at the right of the $u = 0$ line.

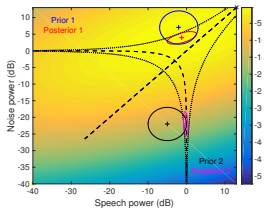

Fig. 6. Plot of two KF posteriors and two priors that are out of the curvy triangle and off the $u = 0$ line. The background is the log-probability of posterior 1.

As seen in Figs. 3-6, the maximum point of the $\alpha = 1$ constraint is at $u = 0$ and $v = -12$ dB. The power constraint $\alpha = 0$ equation, which is in the curvy triangle, is: $v = -2 \log(2\cosh(0.5u))$.

In Figs. 3-6 (and not in the actual algorithm), we assume prior speech and noise independence since the prior distribution is diagonal in Figs. 3-6. Based on Figs. 3-4, when the KF predicted prior distribution is on the $u = 0$ line and has a diagonal covariance matrix (i.e. a diagonal ellipse as a covariance matrix in the $u-v$ rotated domain), the KF posterior distribution must be on the $u = 0$ line and have a diagonal covariance matrix. On the contrary, if the mean of the KF predicted prior distribution is not on the $u = 0$ line, we can say that the mean of the KF posterior distribution is also not on the $u = 0$ line. In the actual algorithm, as presented in Sec. 2.2 and in (3), we do not assume prior speech-noise independence; we track the speech-noise correlation in the log-power spectral domain.

To sum up, we model the inter-frame correlation between adjacent clean speech samples in the log-power spectral domain using a KF prediction step, as explained in Sec. 2.1. We model the correlation between adjacent noise samples using a noise KF prediction step, as presented in Sec. 2.2, and we model the correlation between speech and noise using full covariance matrices. Finally, based on Sec. 2.3, we do a Bayesian KF update step and compute a two-dimensional Gaussian for the speech and noise log-power spectra.

3. IMPLEMENTATION, RESULTS AND EVALUATION

We use acoustic frames of length 32 ms, modulation frames of length 64 ms and a 4 ms acoustic and modulation hop. We use the TIMIT database [27] and the sampling frequency of 16 kHz. We use 40 sentences and 15 noise types from the noise database in [28] at SNR levels of -20 dB to 30 dB. Random segments of noise from the noise

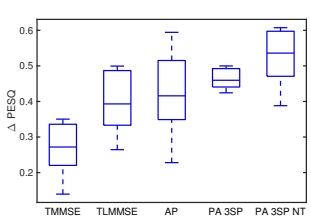

Fig. 7. Boxplot of the ΔPESQ scores for babble noise at 15 dB SNR. The boxplot examines the perceptual speech quality of the presented PA3SP and PA3SPNT algorithms using the median, the first and third quartiles, the minimum and maximum values.

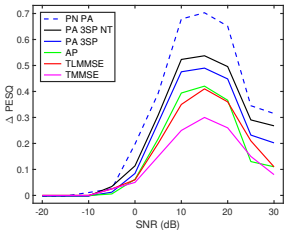

Fig. 8. ΔPESQ for babble noise for the PA3SP, PA3SPNT, the proposed assuming oracle noise PNPA, the proposed KF prediction step with a KF update step assuming additivity in the power domain AP, TMMSE [34] and TLMMSE [33] algorithms.

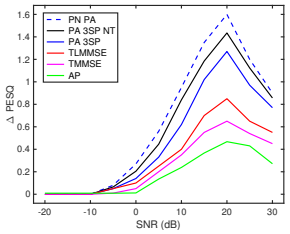

Fig. 9. ΔPESQ for stationary white Gaussian noise.

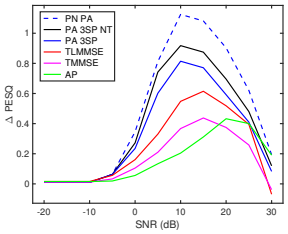

Fig. 10. The ΔPESQ scores for aircraft f16 noise.

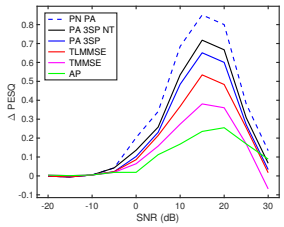

Fig. 11. The ΔPESQ results for non-stationary factory noise.

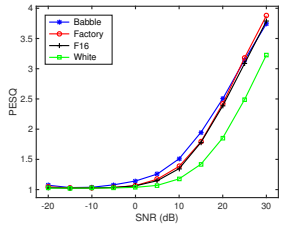

Fig. 12. (Absolute) PESQ scores for noisy speech in Figs. 8-11.

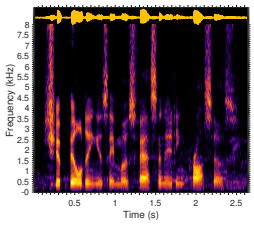

Fig. 13. Plot of the spectrogram of the noisy speech signal. We use 15 dB SNR babble noise.

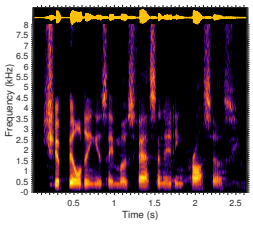

Fig. 14. Plot of the spectrogram of the enhanced speech. We use PA3SP and ΔPESQ $= 0.49$.

signals are utilized. We also use the code from [29], [30] and [31]. We use noise estimation on a frame and frequency basis based on [32]. In Fig. 1, for pre-cleaning, we use the traditional log-MMSE approach [33] and, in Sec. 2.1 and 2.2, we use $p = 2$ and $r = 2$.

We compute the PESQ improvement ΔPESQ in Figs. 7-11. For the KF algorithms, we ignore the initial 0.12 s needed for convergence. We denote our algorithm without noise tracking as PA3SP that refers to the proposed algorithm using three sigma points, as explained in Sec. 2.3. In addition, we denote our algorithm with noise tracking as PA3SPNT that refers to the presented algorithm using three sigma points and noise tracking, as discussed in Sec 2.2.

For comparison purposes, we denote the traditional MMSE approach [34] as TMMSE and the log-MMSE approach [33] as TLMMSE in Figs. 7-11. We implement the TMMSE and TLMMSE algorithms with the same acoustic frame length as our algorithm, which is 32 ms, but with an acoustic frame increment of 16 ms.

In addition, for comparison purposes, in Figs. 7-11, we also use Kalman filtering with a KF update step in another domain: we assume additivity of clean speech and noise in the power domain AP (i.e. $\alpha = 0$), using Gaussian distributions in the log-power domain. We use log-domain speech and noise distributions, as in PA3SP, and we use $\alpha = 0$ in Eq. (5) in Sec. 2.3 in the Bayesian KF update step.

Overall, the alternative possible ways for the KF update step are to: 1) assume speech-noise additivity in the power domain and that speech/noise follow a Gaussian distribution in log-power domain, 2) assume additivity in the amplitude domain (i.e. $\alpha = 1$) and that speech/noise follow a Gaussian distribution in log-power domain, 3) assume additivity in the power domain and that the speech/noise are Gaussian in the power domain, and 4) assume additivity in the amplitude domain and that the speech/noise are Gaussian in the amplitude domain. In [2] and [8], (4) is used and hence $\alpha = 1$ is assumed.

In Figs. 8-11, we also examine the (unrealistic) case when the noise log-power n is perfectly known: PNPA refers to the perfect noise proposed algorithm. Based on the oracle-noise PNPA results, KF-based noise tracking does improve the results of speech enhancement and helps the modified Bayesian KF update step reach a performance that is close to its maximum possible performance.

Figures 7-11 show the ΔPESQ for babble, white Gaussian, f16 and factory noise types. Based on Figs. 7-11, for certain noise types and for a wide SNR range, PA3SP is better than the other algorithms in terms of speech quality with the PESQ metric. In particular, based on Figs. 8-11, the ΔPESQ of PA3SP is better for $-5 < \text{SNR} \leq 30$ dB than the ΔPESQ scores of TMMSE, TLMMSE and AP.

Figures 7-11 also show the ΔPESQ scores of PA3SPNT for babble, white Gaussian, f16 and factory noise types. Noise tracking improves speech enhancement and the speech quality of the enhanced speech signal. The ΔPESQ curves of PA3SPNT approach the ΔPESQ curves of PNPA more than the ΔPESQ curves of PA3SP. Based on Fig. 8, PA3SPNT achieves a ΔPESQ score that is slightly higher than 0.52 at 15 dB SNR at babble noise conditions.

Figure 12 shows the (absolute) PESQ for the noisy speech in Figs. 8-11. Figures 13-14 illustrate spectrograms (power per decade) for noisy and enhanced speech. We use PA3SP and ΔPESQ $= 0.49$ at 15 dB SNR babble noise. In Figs. 13-14, the noise segments (e.g. at 1.5 kHz and 1.25 s) are removed. Finally, it should be noted that the KF algorithms in [2] and [8] are not tested with babble noise.

4. CONCLUSION

In this paper, we presented an algorithm that uses modified Kalman filtering to track speech and noise in the log-power spectral domain. The presented KF update step models the effect of (observation) noise on the clean speech log-power spectrum using phase-sensitive Bayesian estimation. Integration over the phase factor is performed using sigma points. Finally, the results show that the algorithm is better than denoising KF algorithms that assume $\alpha = 0$ or $\alpha = 1$.

5. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Taylor & Francis, Second Edition, ISBN: 978-1-4665-0421-9, pp. 209-234, 2013.

[2] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Communication, 53, 818-829*, 2011.

[3] Y. Wang, "Speech enhancement in the modulation domain," Ph.D. dissertation, Imperial College London, 2015.

[4] Y. Wang and M. Brookes, "Speech enhancement using an MMSE spectral amplitude estimator based on a modulation domain Kalman filter with a Gamma prior," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[5] T. Esch and P. Vary, "Model-based speech enhancement using SNR dependent MMSE estimation," *IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP), pp. 4652-4655*, 2011.

[6] ——, "Speech enhancement using a modified kalman filter based on complex linear prediction and supergaussian priors," *IEEE*, 2008.

[7] T. Esch and P. Vary, "Exploiting temporal correlation of speech and noise magnitudes using a modified Kalman filter for speech enhancement," *Conference on Voice Communication (SprachKommunikation), ITG*, 2008.

[8] Y. Wang and M. Brookes, "Speech enhancement using a modulation domain Kalman filter post-processor with a Gaussian mixture noise model," *Proc. IEEE Intl Conf. Acoustics, Speech and Signal Processing, 7024-7028*, 2014.

[9] J. Li, L. Deng, D. Yu, Y. Gong and A. Acero, "High-performance HMM adaption with joint compensation of additive and convolutive distortions via vector Taylor series," *In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 65-70*, 2007.

[10] ——, "A unified framework of HMM adaption with joint compensation of additive and convolutive distortions," *Comput. Speech Lang. 23 (3), 389-405*, 2009.

[11] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE Trans. Speech Audio Process. 13 (5)*, 2005.

[12] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition, Chapter 6.2: Vector Taylor series, ISBN: 978-0-12-802398-3*. Elsevier, 2016.

[13] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Transactions on Speech and Audio Processing, vol. 12, no. 2*, 2004.

[14] V. S. Leutnant, "Bayesian estimation employing a phase-sensitive observation model for noise and reverberation robust automatic speech recognition," Ph.D. dissertation, Paderborn University, 2015.

[15] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition, Chapter 3.2: Modelling distortions of speech in acoustic environments, and Chapter 3.3: Impact of acoustic distortion on Gaussian modelling, ISBN: 978-0-12-802398-3*. Elsevier, 2016.

[16] T. Yoshioka and D. Sakaue, "Log-normal matrix factorization with application to speech-music separation," *Interspeech, SAPA-SCALE conference*, 2012.

[17] B. D. O. Anderson and J. B. Moore, *Optimal Filtering, ISBN: 978-0486439389*. Dover Publications, 2005.

[18] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing, Chapter 9: Linear predictive analysis of speech signals*. Pearson Education, 2011.

[19] M. H. Hayes, *Statistical digital signal processing and modeling, Chapter 7: Optimum filters*. John Wiley & Sons, 1996.

[20] C. S. J. Doire, M. Brookes, P. A. Naylor et al, "Single-channel online enhancement of speech corrupted by reverberation and noise," *IEEE Transactions on Audio, Speech, and Language Processing, DOI: 10.1109/TASLP.2016.2641904*, 2017.

[21] C. S. J. Doire, "Single-channel enhancement of speech corrupted by reverberation and noise," Ph.D. dissertation, Imperial College London, 2016.

[22] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., R. L. Howell and J. M. Morriss, Eds. McGraw-Hill, Inc., 1991.

[23] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition, Chapter 6.3: Sampling-based methods, ISBN: 978-0-12-802398-3*. Elsevier, 2016.

[24] Y. Cheng and Z. Liu, "Optimized selection of sigma points in the Unscented Kalman filter," *IEEE Transactions on Signal Processing*, 2011.

[25] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceed. IEEE 92 (3), 401-422*, 2004.

[26] ——, "A new extension of the Kalman filter to nonlinear systems," *AeroSense'97, pp. 182-193*, 1997.

[27] J. Garofolo, L. Lamel, W. Fisher et al, "TIMIT accoustic-phonetic continuous speech corpus," *Corpus LDC93S1, Linguistic Data Consortium, Philadelphia*, 1993.

[28] H. Steeneken and F. Geurtsen, "Description of the RSG-10 noise database," *TNO Institute for perception*, 1988.

[29] J. Hartikainen, A. Solin, and S. Sarkka, "Optimal filtering with Kalman filters and smoothers a manual for the MATLAB toolbox EKF/UKF version 1.3," *Department of Biomedical Engineering and Computational Science, Aalto University School of Science, Espoo, Finland*, 2011.

[30] J. H. S. Särkkä and A. Solin, "EKF/UKF: Toolbox for MATLAB v1.3, Bayesian statistical methods: Kalman filtering, Aalto University School of Science," 2007-2016. [Online]. Available: http://becs.aalto.fi/en/research/bayes/ekfukf/

[31] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997-2016.

[32] T. Gerkmann; and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing, Volume: 20, Issue: 4, Pages: 1383-1393*, 2012.

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

[34] ——, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.