Speech Enhancement Using Modulation-Domain Kalman Filtering with Active Speech Level Normalized Log-Spectrum Global Priors

Nikolaos Dionelis and Mike Brookes

Department of Electrical and Electronic Engineering, Imperial College London, UK

***Abstract*—We describe a single-channel speech enhancement algorithm that is based on modulation-domain Kalman filtering that tracks the inter-frame time evolution of the speech log-power spectrum in combination with the long-term average speech log-spectrum. We use offline-trained log-power spectrum global priors incorporated in the Kalman filter prediction and update steps for enhancing noise suppression. In particular, we train and utilize Gaussian mixture model priors for speech in the log-spectral domain that are normalized with respect to the active speech level. The Kalman filter update step uses the log-power spectrum global priors together with the local priors obtained from the Kalman filter prediction step. The log-spectrum Kalman filtering algorithm, which uses the theoretical phase factor distribution and improves the modeling of the modulation features, is evaluated in terms of speech quality. Different algorithm configurations, dependent on whether global priors and/or Kalman filter noise tracking are used, are compared in various noise types.**

I. Introduction

Speech enhancement algorithms can benefit from including a model of the temporal/inter-frame correlation of speech. Based on [1] [2] and on [3], assuming independence between frames is unrealistic and this assumption could be relaxed by imposing temporal structure to the speech model. Inter-frame speech correlation modeling can be performed with a Kalman filter (KF) with a state of low dimension order, based on [4], [5] and [6]. The modulation-domain KF models the short-term time dependencies between successive frames [4] [7].

Existing KF enhancement algorithms that work in the time-frequency domain differ in their choice of the KF state, the KF prediction and the KF update. The KF state can be in the speech amplitude spectral domain [4] [8], the power spectral domain or the log-power spectral domain [9]. Speech spectra are well modelled by Gaussian distributions in the log-power domain (and not so well in other domains) and mean squared errors in the log-power domain are a good measure to use for perceptual speech quality. In addition, the log-power domain is most suitable for infinite-support Gaussian modeling. Regarding the KF prediction, autoregressive (AR) modeling with or without the AR mean can be performed based on the autocorrelation method or the covariance method [10], allowing or not allowing unstable AR poles.

The KF update is affected by the signal model that is used for the addition of speech and noise [11]. If noise and speech are independent then they add in the complex short time Fourier transform (STFT) domain [12] [13]; it may however be analytically simpler to assume that they add either in the power domain or the amplitude domain [4] [8]. The aforementioned alternative possible ways are related to the phase factor, which is the cosine of the phase difference between speech and noise [12] [14]. We can: (a) assume speech and noise additivity in the power spectral domain, using a phase factor equal to zero, or (b) assume additivity in the amplitude spectral domain, using a phase factor equal to unity. In [4] and [8], (b) is used assuming that speech and noise are Gaussian in the amplitude spectral domain. Regarding (b), assuming speech and noise additivity in the amplitude domain results in noise oversubtraction in the region of 0 dB SNR, which may sometimes be perceptually good [15].

Modulation-domain KF algorithms should be able to distinguish between speech and noise. Global speech priors constitute a mechanism that helps in distinguishing between speech and noise. Amongst other technical papers, log-spectrum global priors have been used in denoising nonnegative matrix factorization (NMF) [16] and in logNMF [17]. Speech enhancement can be performed using global priors because a long-term average speech spectrum (LTASS) model exists for speech [18]. By using the long-term average speech log-spectrum, we enhance speech log-spectrum tracking. In this paper, we advance modulation-domain Kalman filtering by utilizing multiple parallel KF updates that use log-spectrum Gaussian Mixture Model (GMM) priors. In [9], we presented a KF-based enhancer that used the log-power spectrum as the KF state and speech-noise additivity in the complex STFT domain as the signal model. In this paper, we extend the KF-based enhancer in [9] to include a GMM speech prior.

II. The speech enhancement algorithm

The flowchart of the algorithm is shown in Fig. 1. The first step is to perform the STFT and then to estimate the active speech level (ASL) [19] [20] and perform ASL normalization. The advantage of ASL normalization is that it permits the use of offline-trained GMM priors that model the distribution of the speech log-spectrum. With the ASL, we have speech models that do not depend on the speech power. The next step is to do Kalman filtering in the log-spectral domain.

In Fig. 1, the blocks in the dotted rectangle constitute the KF. The KF state is the speech log-spectrum and is of

ISBN 978-0-9928626-7-1

dimension $\Re^p$. The KF observation is the noisy speech log-power spectrum y. The algorithm's final step is to keep the first element of the KF state, which is the estimated clean speech in the ASL-normalized log-power spectral domain, transform it to the amplitude domain, denormalize it using the ASL estimate and then reconstuct the clean speech signal using the inverse STFT (ISTFT) and the noisy STFT phase.

A. Notation and the speech-noise signal model

We assume that in the complex STFT domain, the noisy speech is given by $\bar{y}_{\rm d}e^{j\theta} = \bar{s}_{\rm d}e^{j\phi} + \bar{n}_{\rm d}e^{j\psi}$. The amplitudes of the noisy speech, speech and noise are respectively $\bar{y}_{\rm d}$, $\bar{s}_{\rm d}$ and $\bar{n}_{\rm d}$. The subscript "d" denotes that the term is not ASL-normalized. The noisy speech phase is θ, the speech phase is ϕ and the noise phase is ψ. The ASL-normalized spectral amplitudes of the noisy speech, speech and noise are respectively $\bar{y}$, $\bar{s}$ and $\bar{n}$. Using ϵ as the ASL estimate, we have: $\bar{y} = \epsilon^{-0.5}\ \bar{y}_{\rm d}$, $\bar{s} = \epsilon^{-0.5}\ \bar{s}_{\rm d}$ and $\bar{n} = \epsilon^{-0.5}\ \bar{n}_{\rm d}$. The log-powers of the noisy speech, clean speech and noise are respectively denoted by $y = 2\log\bar{y}$, $s = 2\log\bar{s}$ and $n = 2\log\bar{n}$. Within the KF algorithm, we only include the frame index, t, as a subscript in equations involving multiple time frames.

B. The speech KF state and the speech KF prediction

We model the speech time correlation in the speech log-spectrum using the KF prediction step. The speech KF state is the ASL-normalized speech log-power spectrum. Figure 2 shows the speech KF state before and after the KF prediction and update. We utilize the linear KF prediction equations:

$$\mathbf{x}_t = (s_t\ s_{t-1}\ \ldots\ s_{t-p+1})^T \tag{1}$$
$$\mathbf{A}_t = \begin{pmatrix} -\mathbf{a}_t^T \\ \mathbf{I}\ \ \mathbf{0} \end{pmatrix}, \mathbf{Q}_t = \begin{pmatrix} q_0 & \mathbf{0} \\ 0 & \mathbf{0} \end{pmatrix}$$
$$\mathbf{x}_{t+1} = \mathbf{A}_t\mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{x}_t \in \Re^p, \quad \mathbf{A}_t, \mathbf{Q}_t \in \Re^{p\times p}, \quad \mathbf{w}_t \in \Re^p$$

In (1), $\mathbf{x}_t$ is the speech KF state, which contains the current and the past $(p-1)$ speech spectral log-powers. In (1), the KF transition matrix is $\mathbf{A}_t$, the KF transition noise covariance matrix is $\mathbf{Q}_t$ and the KF transition noise is $\mathbf{w}_t$. The KF transition noise $\mathbf{w}_t$ is Gaussian, zero-mean and has $\mathbf{Q}_t$ as its covariance matrix. The KF transition matrix $\mathbf{A}_t$ is from AR modeling; AR(p) modeling defines the dimensions of the matrices in the KF prediction. The speech AR parameters are $\mathbf{a}_t \in \Re^p$ and q_0 is the AR modeling error variance.

We use a time-varying KF: the transition matrix $\mathbf{A}_t$ and the transition noise covariance matrix $\mathbf{Q}_t$ depend on AR modeling using the covariance method [10] on the pre-cleaned modulation frame, estimating both the AR coefficients and the AR mean of clean speech. The AR mean is the average clean speech log-power that is estimated as an AR parameter.

In (1), we define how the speech KF state $\mathbf{x}_t$ changes in the KF prediction. The speech KF state consists of a speech KF state mean and a speech KF state covariance matrix. Considering the linear KF prediction equations and using (1), both the speech KF state mean and the speech KF state covariance matrix are updated in the KF prediction step. In Fig. 2, the speech KF state mean is denoted by $\boldsymbol{\mu}_t$ and the speech KF state covariance matrix is denoted by $\mathbf{P}_t$.

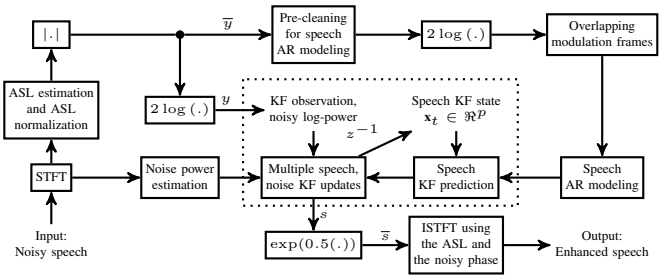

Figure 1. The flowchart diagram of the algorithm. The term z^{-1} refers to one-frame delay. The blocks in the dotted rectangle constitute the KF.

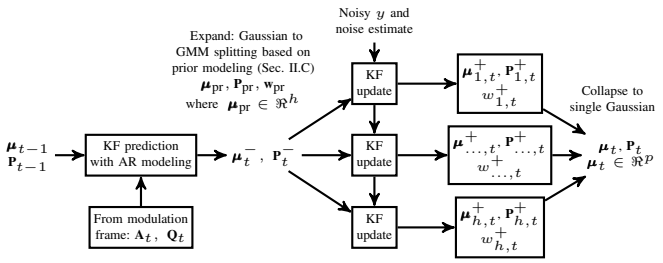

Figure 2. The speech KF is shown. We focus on the speech KF state. We expand to h weighted Gaussians based on our Gaussian splitting algorithm using offline-trained log-spectrum global priors, as described in Sec. II.C.

C. The log-spectrum global speech priors

Based on Figs. 1-2, we perform multiple speech-noise KF updates due to using a GMM of h mixtures as global speech priors. We use global priors together with the KF-based local priors. We use a Gaussian splitting algorithm that is based on ASL-normalized offline-trained priors. We multiply the current element of the decorrelated KF state with the global priors. Decorrelation and correlation of the KF state are used to preserve the KF prediction inter-frame modeling. We first decompose the speech KF state covariance matrix $\mathbf{P}$ as:

$$\mathbf{P} = \begin{pmatrix} g_0 & \mathbf{g}^T \\ \mathbf{g} & \mathbf{G} \end{pmatrix} \tag{2}$$

where g_0 is the variance of the current element of the speech KF state. We define the linear transformation matrix $\mathbf{B}$ by [5]:

$$\mathbf{B} = \begin{pmatrix} 1 & \mathbf{0}^T \\ -g_0^{-1}\mathbf{g} & \mathbf{I} \end{pmatrix} \tag{3}$$

The next step is to compute the linearly transformed speech KF state $\mathbf{B}\mathbf{x}_t$ with mean $\mathbf{B}\boldsymbol{\mu}$ and with covariance matrix [5]:

$$\mathbf{B}\ \mathbf{P}\ \mathbf{B}^T = \begin{pmatrix} g_0 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{G} - g_0^{-1}\mathbf{g}\mathbf{g}^T \end{pmatrix} \tag{4}$$

In (4), g_0 is preserved. After the multiple parallel KF updates, we correlate the KF state by using $\mathbf{B}^{-1}$ and the

ISBN 978-0-9928626-7-1 © EURASIP 2017

inverse of the linear transformation in (4). We use speech GMM priors that are multiplied with the current element of the decorrelated speech KF state after the KF prediction. The decorrelated speech KF state is $\mathbf{B}\mathbf{x}_t$ so that the current speech log-power s_t is uncorrelated with $(s_{t-1} \ ... \ s_{t-p+1})^T$.

In Fig. 2, we compute the posterior weights $w_{i,t}^+$ for $i \in [1, h]$ after each of the multiple KF updates. Finding $w_{i,t}^+$ involves the use of the GMM KF update [21], which in turn involves the use of the nonlinear KF observation model.

D. The phase-factor-sensitive modified KF update

The KF update estimates the posterior of the speech and noise log-powers given the noisy log-power. The KF update is described in more detail in [9]. The KF update considers the Gaussian speech and noise priors from the KF prediction, the distribution of the STFT phase difference between speech and noise using the phase factor $\alpha = \cos(\phi - \psi)$ [12] [13]:

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

From (5): $\alpha = 0.5\exp(y - 0.5(s+n)) - \cosh(0.5u)$. We use $u = n - s$ and $y = 0.5(s+n) + \log(2(\alpha + \cosh(0.5u)))$.

We do the variable transformation $(s, n, \alpha) \Rightarrow (u, y, \alpha)$. The Jacobian determinant is $\Delta = 1$. Now, the posterior is:

$$\begin{aligned} p(u, \alpha \,|\, y) &= \left.\frac{p(u, \alpha, y)}{p(y)}\right|_y \propto p(u, \alpha, y)|_y \\ &\propto \left.\left(p(s,n)\ p(\alpha)\ |\Delta|^{-1}\right)\right|_y \propto p(\alpha) \\ &\times \mathcal{N}\left(\begin{pmatrix} y - \log(2(\alpha + \cosh(0.5u))) - 0.5u \\ y - \log(2(\alpha + \cosh(0.5u))) + 0.5u \end{pmatrix}; \mathbf{m}, \mathbf{S}\right) \end{aligned} \tag{6}$$

where $p(\alpha) = \left(\pi\sqrt{1-\alpha^2}\right)^{-1}$ for $-1 < \alpha < 1$ and zero otherwise. Here, we assume that ψ is uniform $\psi \sim U(-\pi, \pi)$ and thus $(\phi - \psi) \sim U(-\pi, \pi)$ [14]. In (6), for $p(s, n)$, we use a Gaussian with mean $\mathbf{m}$ and covariance matrix $\mathbf{S}$. As in [9], we find the moments of the posterior (s, n) using (7) for $0 \le a + b \le 2$. We use $s = s(u, y, \alpha)$ and $n = n(u, y, \alpha)$.

$$\begin{aligned} E\{s^a n^b \mid y\} &= \int_{\alpha=-1}^{1}\int_{u=-\infty}^{\infty} s^a n^b\ p(u, \alpha \mid y)\ du\ d\alpha \\ &= \frac{1}{|\Delta|\ p(y)} \int_{\alpha=-1}^{1} p(\alpha) \int_{u=-\infty}^{\infty} s^a n^b\ p(s, n)\ du\ d\alpha \end{aligned} \tag{7}$$

In (7), the integration over u is performed with truncated Gaussians and straight line segments, obtaining a closed-form solution. The integration over α is done using R sigma points, as in [9], utilizing the Unscented transform [22] [23]. In (7), $E\{\alpha^z\}$ is needed for the integration over α with sigma points: $E\{\alpha^z\} = 2^{-z} z!\left((0.5z)!\right)^{-2}$ for even z and zero otherwise.

III. Noise tracking and the speech-noise KF

We now present the noise KF state, the noise KF prediction [24] and the speech-noise KF prediction. With noise tracking, the (s, n) priors are correlated and the KF state $\in \Re^{p+q}$ is the speech KF state $\in \Re^p$ and the noise KF state $\in \Re^q$.

We do noise tracking based on AR(q) modeling and on the estimated SNR in the modulation frame [6] [9]. After the noise KF prediction, we decorrelate the noise 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 [25] [26]. As in (1) that describes the speech KF prediction, for the noise, (n), KF prediction:

$$\begin{aligned} &\mathbf{x}_t^{(\mathrm{n})} = (n_t \ n_{t-1} \ ... \ n_{t-q+1})^T \in \Re^q \\ &\mathbf{A}_t^{(\mathrm{n})} = \begin{pmatrix} -\mathbf{a}_t^{(\mathrm{n})T} \\ \mathbf{I} \ \ \mathbf{0} \end{pmatrix} \in \Re^{q\times q}, \quad \mathbf{Q}_t^{(\mathrm{n})} = \begin{pmatrix} q_0^{(\mathrm{n})} & \mathbf{0} \\ 0 & \mathbf{0} \end{pmatrix} \in \Re^{q\times q} \\ &\mathbf{x}_{t+1}^{(\mathrm{n})} = \mathbf{A}_t^{(\mathrm{n})}\mathbf{x}_t^{(\mathrm{n})} + \mathbf{w}_t^{(\mathrm{n})}, \quad \mathbf{w}_t^{(\mathrm{n})} \in \Re^q \end{aligned} \tag{8}$$

The joint, (j), speech-noise KF state $\mathbf{z}_t$ is defined in (9). We use full covariance matrices due to the KF update in Sec. II.D.

$$\begin{aligned} &\mathbf{z}_t = \left(\mathbf{x}_t^T \ \mathbf{x}_t^{(\mathrm{n})T}\right)^T \in \Re^{p+q}, \quad \mathbf{A}_t^{(\mathrm{j})} = \begin{pmatrix} \mathbf{A}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_t^{(\mathrm{n})} \end{pmatrix} \\ &\mathbf{Q}_t^{(\mathrm{j})} = \begin{pmatrix} \mathbf{Q}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_t^{(\mathrm{n})} \end{pmatrix}, \quad \mathbf{A}_t^{(\mathrm{j})}, \mathbf{Q}_t^{(\mathrm{j})} \in \Re^{(p+q)\times(p+q)} \\ &\mathbf{z}_{t+1} = \mathbf{A}_t^{(\mathrm{j})}\mathbf{z}_t + \mathbf{w}_t^{(\mathrm{j})}, \quad \mathbf{w}_t^{(\mathrm{j})} \in \Re^{p+q} \end{aligned} \tag{9}$$

IV. Implementation, Results and Evaluation

We use acoustic frames of length 32 ms, modulation frames of length 32 ms or 64 ms and a 4 ms acoustic and modulation frame increment. We use the TIMIT database [27] sampled at 16 kHz. For the training of the global speech priors, we use 250 sentences and for testing, we use 40 sentences. We use noise types from the noise database in [28] at SNR levels from -20 dB to 30 dB. Random segments of noise from the noise signals are used [29]. The external noise estimation is based on [30] [29]. For pre-cleaning in Fig. 1, we use the traditional log-MMSE approach [31] [29]. In Secs. II.B and III, we use $p = 2$ and $q = 2$. In Secs. II.C-D, $h = 4$ and $R = 3$.

For evaluation purposes, we compare the results with and without global speech priors, and with and without noise KF tracking. We consider alternative configurations of the algorithm in Figs. 1-2. Table I shows the Bark Spectral Distortion (BSD) [32] for babble noise at 15 dB SNR. We compute the BSD using no voice activity detection. In Table I, the BSD of the noisy speech signal is 2.64×10^{-2} dB.

Table I
BSD ($\times 10^{-2}$ dB) for babble noise at 15 dB SNR.

ST	SMST	NTST	GPST	SMGPST
0.98	0.95	0.92	0.93	0.92
NTGPST	EEST	NTEEST	SMEEST	NTSMEEST
0.90	0.91	0.90	0.90	0.88

ST = Speech tracking: the KF tracks the clean speech log-spectrum.
SM = Smaller modulation frame: 32 ms instead of 64 ms.
NT = Noise tracking: the KF tracks the noise log-spectrum.
GP = Global priors: we use Fig. 2 and log-spectrum speech priors.
EE = Early expanding using the log-spectrum speech priors before the KF prediction. EE assumes GP and EE changes Fig. 2.

ISBN 978-0-9928626-7-1 © EURASIP 2017

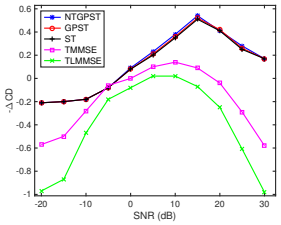

Figure 3. $-\Delta$CD for babble noise.

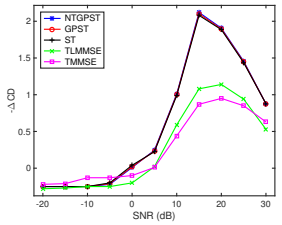

Figure 4. $-\Delta$CD for white noise.

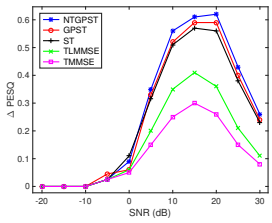

Figure 7. ΔPESQ for babble noise.

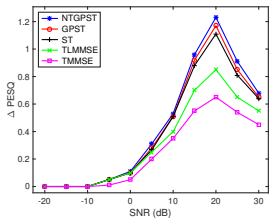

Figure 8. ΔPESQ for white noise.

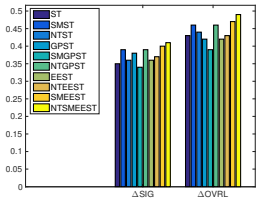

Figure 5. ΔSIG and ΔOVRL for babble noise at 15 dB SNR.

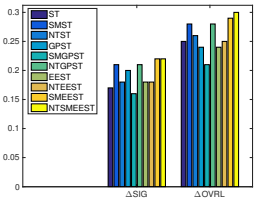

Figure 6. ΔSIG and ΔOVRL for babble noise at 5 dB SNR.

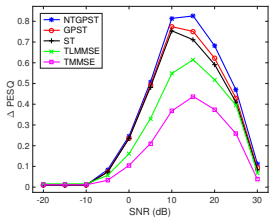

Figure 9. Plot of the ΔPESQ scores for aircraft f16 noise.

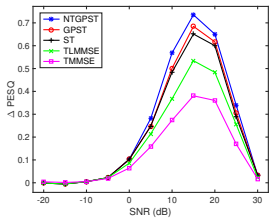

Figure 10. Plot of the ΔPESQ for non-stationary factory noise.

Based on Table I, the ST algorithm that does not perform KF noise tracking has a higher BSD than the global priors ST (GPST). This means that the offline-trained priors aid speech tracking; using global speech priors reduces the BSD.

In Table I, we consider early expanding (EE). Figure 2 does late expanding since the global speech priors are used after the KF prediction. On the contrary, with EE, the global priors are used before the KF prediction: the Gaussian-GMM multiplication is performed before the KF prediction. Comparing GPST with EEST in Table I, we note that EE reduces the BSD.

In Table I, using smaller modulation frames (SM) reduces the BSD. The tradeoff is between noisier AR modeling and a modulation frame that is more concentrated in time.

We now use noise tracking and global priors (NTGP). In Table I, we observe a decreasing error from ST to GPST and to NTGPST. With the global priors, as presented in Fig. 2, the BSD error is 0.90×10^{-2} dB at 15 dB SNR babble noise.

We now examine babble noise at 5 dB SNR. Like Table I, Table II shows the BSD. The same algorithm notation as in Table I is used. We see a decreasing error from ST to GPST and to NTGPST. The noisy speech BSD is 1.83×10^{-1} dB.

Table II
BSD ($\times 10^{-1}$ dB) FOR BABBLE NOISE AT 5 dB SNR.

ST	SMST	NTST	GPST	SMGPST
0.67	0.65	0.62	0.64	0.61
NTGPST	EEST	NTEEST	SMEEST	NTSMEEST
0.59	0.62	0.60	0.58	0.58

Figures 3-4 show the Cepstrum Distance (CD) as a speech quality metric [33]. The cepstrum is directly related to the minimization of the log-power error that we want to achieve with log-spectrum Kalman filtering. Figures 3-4 depict the negative CD improvement $-\Delta$CD of the algorithms for babble and white noises. The $-\Delta$CD are positive at high SNRs.

Tables I-II examine the alternative configurations of the proposed algorithm for specific SNRs. On the contrary, Figs. 3-4 compare the alternative configurations of the proposed algorithm with traditional speech ehnhancement techniques in the SNR range of -20 dB to 30 dB. For comparison purposes, we denote the traditional MMSE approach [34] as TMMSE and the traditional log-MMSE approach [31] as TLMMSE.

We use the speech distortion SIG, noise distortion BAK and overall quality OVRL metrics from [35] [15], which are in a scale of 1 to 5 where 5 indicates excellent speech quality. Figures 5-6 illustrate the ΔSIG and ΔOVRL for babble noise at 15 dB and 5 dB SNR. Considering a specific case, in 15 dB babble noise, ST has the ΔOVRL score of 0.53.

In Figs. 7-10, we use the PESQ speech quality metric for babble, white, aircraft f16 and factory noises. In Figs. 7-10, the presented KF-based algorithms are better than the TLMMSE and TMMSE. We observe that in the SNR range of 0 dB to 30 dB, the presented KF algorithms outperform the traditional noise suppression techniques. We also observe that the best performance of the presented algorithm is when both noise KF tracking and global speech priors are used. As in Figs. 7-10, the ST algorithm is also evaluated in [9] with PESQ.

V. Conclusion

In this paper, we present a single-channel speech enhancement algorithm that is based on modulation-domain Kalman filtering that tracks the time evolution of the speech log-power spectrum in every frequency using the long-term average speech log-spectrum. The noise suppression algorithm applies a KF that uses offline-trained log-spectrum priors that are normalized with respect to the active speech level. Denoising is performed with active speech level normalized log-spectrum global priors, by training and utilizing Gaussian mixture models. The KF update uses the phase factor between speech and noise. The KF algorithm is evaluated in terms of speech quality and different algorithm configurations are compared.

ISBN 978-0-9928626-7-1 © EURASIP 2017

REFERENCES

[1] D. Liang, M. D. Hoffman and G. J. Mysore, "Speech dereverberation using a learned speech model," *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2015.

[2] N. Boulanger-Lewandowski, G. J. Mysore and M. Hoffman, "Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation," *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014.

[3] I. Andrianakis and P. R. White, "On the application of Markov Random Fields to speech enhancement," *International Conference on Mathematical Signal Processing (IMA)*, 2006.

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

[5] 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.

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

[7] S. So, K. K. Wojcicki and K. K. Paliwal, "Single-channel speech enhancement using Kalman filtering in the modulation domain," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2010.

[8] Y. Wang and M. Brookes, "Speech enhancement using a modulation domain Kalman filter post-processor with a Gaussian mixture noise model," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[9] N. Dionelis and M. Brookes, "Modulation-domain speech enhancement using a Kalman filter with a Bayesian update of speech and noise in the log-spectral domain," *IEEE International Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2017.

[10] 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.

[11] 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.

[12] 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 Trans. on Speech and Audio Processing, vol. 12, no. 2*, 2004.

[13] ——, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. on Speech and Audio Processing, vol. 12, no. 3*, 2004.

[14] V. Leutnant and R. Haeb-Umbach, "An analytic derivation of a phase-sensitive observation model for noise-robust speech recognition," *Conference of the International Speech Communication Association (INTERSPEECH), pp. 2395-2398*, 2009.

[15] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Taylor & Francis, Second Edition, 2013.

[16] K. W. Wilson, B. Raj, P. Smaragdis and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," *IEEE International Conference on Audio and Speech Signal Processing (ICASSP)*, 2008.

[17] T. Yoshioka and D. Sakaue, "Log-normal matrix factorization with application to speech-music separation," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[18] D. Byrne et al, "An international comparison of long-term average speech spectra," *The Journal of the Acoustical Society of America, vol. 96, no. 4*, 1994.

[19] N. Dionelis and M. Brookes, "Active speech level estimation in noisy signals with quadrature noise suppression," *European Signal Processing Conference (EUSIPCO)*, 2016.

[20] S. Gonzalez and M. Brookes, "Speech active level estimation in noisy conditions." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[21] J. H. Kotecha and P. M. Djuric, "Gaussian sum particle filtering," *IEEE Trans. on Signal Processing, vol. 51, no. 10*, 2003.

[22] 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.

[23] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE, vol. 92, no. 3*, 2004.

[24] B. Raj, R. Singh and R. Stern, "On tracking noise with linear dynamical system models," *Conference of the International Speech Communication Association (INTERSPEECH), pp. 965-968*, 2004.

[25] C. S. J. Doire, M. Brookes, P. A. Naylor et al, "Single-channel online enhancement of speech corrupted by reverberation and noise," *IEEE/ACM Trans. on Audio, Speech and Language Processing, vol. 25, no.3*, 2017.

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

[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] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997-2017.

[30] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Language Processing, vol. 20, no. 4*, 2012.

[31] 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. 33, no. 2, 1985.

[32] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, June 1992.

[33] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low bit-rate speech coding systems," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 262–273, 1988.

[34] 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. 32, no. 6, pp. 1109–1121, 1984.

[35] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2006.

ISBN 978-0-9928626-7-1 © EURASIP 2017