

A Kalman filtering algorithm with joint metrics-based tuning for single-channel speech enhancement

Aidan E.W. George, Stephen So, Ratna Ghosh, Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering,
Griffith University, Brisbane, QLD, Australia, 4111.

{a.george, s.so, k.paliwal}@griffith.edu.au, rg@iee.jusl.ac.in

Abstract

In this paper, we present an iterative Kalman filtering algorithm that exhibits better speech enhancement by jointly utilising robustness and sensitivity metrics. Typically, poor model parameter estimates lead to a biased Kalman filter gain, which results in innovation noise ‘leaking’ into the output. In the proposed algorithm, the Kalman filter gain is dynamically tuned based on a varying operating point of balanced robustness and sensitivity. Speech enhancement experiments showed the proposed Kalman filtering algorithm to produce higher quality speech than conventional methods using objective and subjective measures.

Index Terms: Speech enhancement, Kalman filtering, noise reduction

1. Introduction

The role of speech enhancement is to reduce the level of undesirable background noise in digitally-recorded speech in order to improve its quality and intelligibility. Several speech enhancement algorithms have been reported in the literature that have had varying degrees of success in enhancing speech but many of them suffer from problems with residual noise, such as the musical noise typically present in Wiener filtering and spectral subtraction algorithms [1]. The Kalman filter was first applied to speech enhancement by Paliwal and Basu [2], and since then has been investigated in the literature both in the time-domain (e.g. [3, 4, 5]) and the modulation-domain [6].

The Kalman filter is an unbiased and linear minimum-mean-squared-error (MMSE) estimator [7] that estimates the clean state vector of speech samples by using a weighted combination of predictions from a speech production model and noise-corrupted speech measurements. The autoregressive (AR) model of speech is commonly used with Kalman filtering to provide the predicted component. While performing remarkably well in the *oracle* case, where AR parameters from the clean speech were available [2], the Kalman filter exhibits poor enhancement performance in practice (non-oracle), where only the noise-corrupted speech is available. This is because the presence of noise leads to bias in the AR parameter estimates. This has a detrimental effect on the enhancement ability in the regions where speech is absent, since the AR estimation bias offsets the *Kalman filter gain*, which regulates how much of the (noisy) innovation signal is used to correct the AR model prediction [5]. This Kalman filter gain offset in the speech-absent regions results in noise from the innovation signal ‘leaking’ into the output.

In this paper, we propose a new Kalman filtering algorithm that reduces the detrimental effects of poor AR parameter estimates on enhancement performance by jointly utilising two metrics (robustness and sensitivity) [8] that are computed in real-time. The algorithm is iterative in nature. In the initial iteration, the Kalman filter is operated in a constrained mode, where the sensitivity and robustness metrics are balanced. Then, in the subsequent iteration, the AR parameters are estimated from the pre-processed speech and then used in a delayed Kalman filter [2]. Experiments were performed on

speech that was corrupted with white Gaussian noise (WGN). The experimental results presented in this paper show that the proposed algorithm exhibited a large improvement in performance over the normal (non-oracle) iterative Kalman filter¹.

2. Kalman filter-based speech enhancement

The additive noise model generally assumed in the problem of speech enhancement is:

$$y(n) = x(n) + v(n) \quad (1)$$

where $x(n)$ is the clean speech, $v(n)$ is a white Gaussian noise (with a variance of σ_v^2), and $y(n)$ is the noise-corrupted speech. The speech and noise signals are assumed to be zero-mean and uncorrelated with each other. In the Kalman filter, a p th order autoregressive (AR) model is used to represent speech production, whose parameters $\{a_k; k = 1, 2, \dots, p\}$ and σ_w^2 , represent the AR coefficients and excitation noise variance, respectively [5].

The Kalman filter recursively computes an *a posteriori* state vector estimate $\hat{\mathbf{x}}(n|n)$ at time n , when given the noisy speech measurement $y(n)$ and the *a priori* state vector estimate $\hat{\mathbf{x}}(n|n-1)$. For a detailed description of each variable, the reader is referred to [5, 7]:

$$\mathbf{P}(n|n-1) = \mathbf{A}\mathbf{P}(n-1|n-1)\mathbf{A}^T + \sigma_w^2 \mathbf{d}\mathbf{d}^T \quad (2)$$

$$\mathbf{K}(n) = \mathbf{P}(n|n-1)\mathbf{c} \left[\sigma_v^2 + \mathbf{c}^T \mathbf{P}(n|n-1)\mathbf{c} \right]^{-1} \quad (3)$$

$$\mathbf{P}(n|n) = [\mathbf{I} - \mathbf{K}(n)\mathbf{c}^T] \mathbf{P}(n|n-1) \quad (4)$$

$$\hat{\mathbf{x}}(n|n-1) = \mathbf{A}\hat{\mathbf{x}}(n-1|n-1) \quad (5)$$

$$\hat{\mathbf{x}}(n|n) = \hat{\mathbf{x}}(n|n-1) + \mathbf{K}(n)[y(n) - \mathbf{c}^T \hat{\mathbf{x}}(n|n-1)] \quad (6)$$

In conventional Kalman filtering, the enhanced speech signal at the present time n is given by²:

$$\hat{x}(n|n) = \mathbf{c}^T \hat{\mathbf{x}}(n|n) \quad (7)$$

This is equivalent to taking the first scalar component of the state vector $\hat{\mathbf{x}}(n|n)$. Therefore, it is possible to re-write some of the Kalman recursion equations in scalar form.

$$\hat{x}(n|n) = \hat{x}(n|n-1) + K(n)[y(n) - \hat{x}(n|n-1)] \quad (8)$$

$$= [1 - K(n)]\hat{x}(n|n-1) + K(n)y(n) \quad (9)$$

where $\hat{x}(n|n-1)$ and $K(n) = \mathbf{c}^T \mathbf{K}(n)$ are the first scalar components of the *a priori* state vectors and Kalman gain vector, respectively. It can be seen that the scalar Kalman filter gain adjusts the relative proportions of the noisy observation sample

¹The MATLAB code and sample speech output files are available at http://tiny.cc/speech_enhancement

²We used **bold** variables to denote matrix/vector quantities, as opposed to unbolded variables for scalar quantities.

$y(n)$ and *a priori* sample $\hat{x}(n|n-1)$.

We can re-write Eq. (3) in terms of scalar quantities [9]:

$$K(n) = \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \sigma_v^2} \quad (10)$$

where $\alpha^2(n) = \mathbf{c}^T \mathbf{A} \mathbf{P}(n-1|n-1) \mathbf{A}^T \mathbf{c}$ represents the contribution of the *a posteriori* mean squared error from the previous time step $n-1$, to the total *a priori* mean squared error of the speech model prediction.

Equation (10) provides further insight into the operation of the Kalman filter when used in speech enhancement. When there is no corrupting noise (i.e. $\sigma_v^2 = 0$), the scalar Kalman filter gain becomes unity and therefore, when substituted into Eq. (9), the enhanced speech sample is formed from the observed speech sample $y(n)$ only. On the other hand, when the corrupting noise levels are much higher than the *a priori* mean squared error (i.e. $\sigma_v^2 \gg \alpha^2(n) + \sigma_w^2$), the Kalman filter will favour the predicted speech sample computed from the speech production model.

It is known that noise in the speech will add bias to the AR parameter estimates, which in turn will have a degrading effect on the enhancement performance of the Kalman filter. In this study, we focus our attention on the biased excitation variance estimate, $\tilde{\sigma}_w^2$. For corrupting noise that is white and Gaussian, it can be generally assumed³ that $\tilde{\sigma}_w^2 \approx \sigma_w^2 + \sigma_v^2$. Therefore, after substituting this into Eq. (10), the scalar Kalman filter gain will also become biased:

$$\tilde{K}(n) = \frac{\alpha^2(n) + \tilde{\sigma}_w^2}{\alpha^2(n) + \tilde{\sigma}_w^2 + \sigma_v^2} \quad (11)$$

$$\approx \frac{\alpha^2(n) + \sigma_w^2 + \sigma_v^2}{\alpha^2(n) + \sigma_w^2 + 2\sigma_v^2} \quad (12)$$

The negative effect of the biased scalar Kalman filter gain manifests itself mostly in the silent pauses between spoken words. Since speech is not present in the silent pauses (i.e. $\sigma_w^2 = \alpha^2(n) = 0$), the biased Kalman filter gain will fluctuate around 0.5. According to Eq. (9), this means that half of the corrupting noise $y(n)$ is passed through to the output.

In this paper, we will consider a new algorithm for reducing the effect of biased estimates on the scalar Kalman filter gain. Previous studies have investigated the use of iterative techniques [3] to reduce estimation bias, where the AR coefficients and excitation variance are re-estimated from filtered speech and then are used in the next iteration. In practice, iterative Kalman filters often suffer from speech distortion and musical noise. However, it has been shown that the initial iteration is important and that better parameter estimates in this iteration will generally result in better performance [5]. In this study, we aim to improve the iterative Kalman filter by quantitatively monitoring the effects of estimation bias using two performance metrics and then tuning the Kalman filter gain in the first iteration.

3. Proposed Kalman filtering algorithm

3.1. Joint robustness and sensitivity metrics

Robustness relates to the ability of the Kalman filter to mitigate uncertainty in its dynamic model parameters. A performance metric for measuring the level of robustness in the Kalman filter was proposed in the instrumentation literature [8]. The robustness metric $J_2(n)$ can be rewritten in terms of scalar quantities as:

$$J_2(n) = \frac{\sigma_w^2}{\alpha^2(n) + \sigma_w^2} \quad (13)$$

³This assumes that the bias in the AR coefficients a_k is negligible.

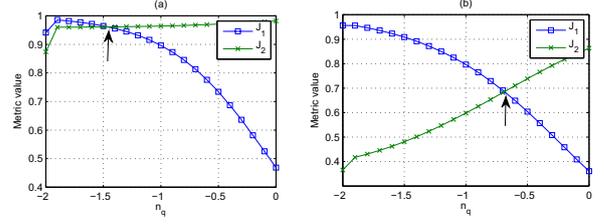


Figure 1: Plot of average Kalman filter sensitivity and robustness metrics (J_1 and J_2) over a frame for varying speech excitation variance $\hat{\sigma}_w^2 = (10^{n_q})\sigma_w^2$ (compromise point where $J_1 = J_2$ shown by the arrow) in a: (a) speech-absent region (only noise); and (b) voiced-speech region.

It can be observed that the speech excitation variance σ_w^2 plays an important role in determining how robust the Kalman filter is. In other words, a large σ_w^2 indicates an unreliable speech model, such as that for voiced speech, which has strong harmonic structure that cannot be predicted by a low-order AR model.

The *sensitivity* metric $J_1(n)$, which quantifies the ability of the Kalman filter to respond to dynamic changes in the input speech (and accordingly, changes in the speech model parameters) in order to mitigate the effects of measurement noise, can be expressed in scalar quantities as:

$$J_1(n) = \frac{\sigma_v^2}{\alpha^2(n) + \sigma_v^2 + \sigma_w^2} \quad (14)$$

The sensitivity metric is dependent on the measurement noise variance σ_v^2 . In the case where the measurement signal is too highly corrupted with noise (i.e. $\sigma_v^2 \gg [\sigma_w^2 + \alpha^2(n)]$), the Kalman filter becomes more reliant on its speech model.

The excitation noise variance σ_w^2 , which measures the uncertainty of the speech model, appears in both equations for $J_1(n)$ and $J_2(n)$. Changes in this variance term will invariably lead to variations in the metrics [8]. Figure 1 shows a plot of the two metrics (averaged over a frame) as σ_w^2 is varied, i.e. $\hat{\sigma}_w^2 = (10^{n_q})\sigma_w^2$, within regions where there is no speech [Figure 1(a)] and where there is voiced speech [Figure 1(b)]. We can see that the compromise point of *balanced* robustness and sensitivity varies a lot between frames of only noise and frames of voiced speech. Similarly, the values of the metric averages (J_1 and J_2) at this compromise point vary between 0 and 1. The value of $\hat{\sigma}_{wc}^2$ (at this compromise point) can be derived by equating Eqs. (13) and (14):

$$\frac{\sigma_v^2}{\alpha^2(n) + \sigma_v^2 + \hat{\sigma}_{wc}^2} = \frac{\hat{\sigma}_{wc}^2}{\alpha^2(n) + \hat{\sigma}_{wc}^2}$$

Solving for $\hat{\sigma}_{wc}^2$ (and keeping only the positive root), we can compute the excitation variance when the sensitivity and robustness metrics are balanced:

$$\hat{\sigma}_{wc}^2 = \frac{\alpha(n)\sqrt{\alpha^2(n) + 4\sigma_v^2} - \alpha^2(n)}{2} \quad (15)$$

A recent method that used only the robustness metric to tune the Kalman filter was presented in [9]. In this method, the Kalman filter gain was tuned using the value of $J_2(n)$ in order to reduce the effects of bias:

$$K'(n) = [1 - J_2(n)]K(n) \quad (16)$$

One of the problems with this method was that there was over-suppression of the Kalman filter gain even in the speech regions, which introduced speech distortion. The value of $J_2(n)$ (seen in Figure 1 at $n_q = 0$) can be seen to not vary much between noise-only and speech-present frames. In our proposed method,

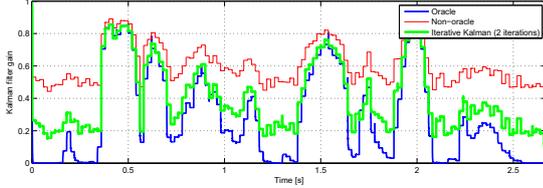


Figure 2: Plot of the scalar Kalman filter gain for oracle and non-oracle modes compared with that of the conventional iterative Kalman filter (with two iterations) [3]. Utterance was sp10 (“The sky that morning was clear and bright blue”) corrupted with WGN at 5 dB SNR.

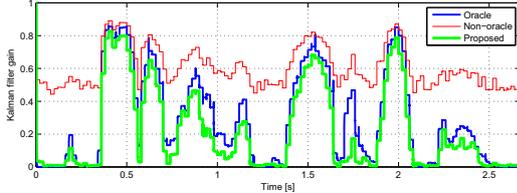


Figure 3: Plot of the scalar Kalman filter gain for oracle and non-oracle modes compared with that of the proposed iterative Kalman filter (with two iterations). Utterance was sp10 (“The sky that morning was clear and bright blue”) corrupted with WGN at 5 dB SNR.

instead of using $J_2(n)$ alone in the Kalman filter gain tuning, we utilise $J_{2c}(n)$ ($J_2(n)$ at the compromise point), which combines the effect of the sensitivity metric $J_1(n)$, as was seen in Figure 1. This enables the tuning to better handle both speech-absent and speech-present frames.

We can summarise the proposed iterative Kalman filtering algorithm. For each frame:

- Step 1:** In the *first iteration*, compute the value of $\hat{\sigma}_{wc}^2$ using Eq. (15);
- Step 2:** Substitute into Eq. (13) to compute the robustness metric at the compromise point $J_{2c}(n)$;
- Step 3:** After computing Eq. (13), adjust the Kalman filter gain as in Eq. (16) using $K'(n) = [1 - J_{2c}(n)]K(n)$;
- Step 4:** In the *second iteration*, estimate the AR parameters from the enhanced speech of the first pass and filter speech using the delayed Kalman filtering algorithm [2] with no Kalman filter gain modification.

3.2. Comparison of the proposed algorithm with the conventional iterative Kalman filter

In this section, we will compare the performance of the proposed iterative Kalman filtering algorithm with the conventional iterative algorithm of Gibson, et al. [3]. For the basis of comparison, both methods use two iterations, with the delayed version of the Kalman filter in the second iteration. Figures 2 and 3 show the scalar Kalman filter gain trajectories for the conventional and proposed iterative Kalman filtering algorithms, respectively. We can see that in the proposed algorithm, the gain is better suppressed in the silence regions, when compared with the conventional one. In the speech regions, the gain in the proposed algorithm is generally similar to what is seen in the oracle case.

4. Experimental setup

The NOIZEUS speech corpus [1] was used in our enhancement experiments, which is composed of 30 phonetically balanced sentences belonging to six speakers. The sampling frequency

was 8 kHz. For our objective experiments, we generated a set of stimuli that has been corrupted by WGN at different SNR levels. The segmental signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) [1] were used to evaluate the following treatment types.

1. Speech corrupted with white Gaussian noise (**No enhancement, noisy**);
2. Kalman filter (delayed) with AR parameters estimated from clean speech [2] (**Kalman oracle**);
3. Kalman filter (iterative) [3] with two iterations (**Kalman iterative**);
4. Proposed Kalman filter with two iterations (**Kalman proposed**); and
5. Minimum mean squared error short-time spectral amplitude estimator [10] (**MMSE-STSA**).

In order to determine the subjective quality of the proposed method in comparison with the other speech enhancement algorithms, a blind AB listening test [4] was performed. Pairs of stimuli were played back to 11 English-speaking listeners, who were then asked to make a subjective preference for each pair. The speech utterance ‘She had a smart way of wearing clothes’ was corrupted by white Gaussian noise at an SNR of 10 dB. The total number of pair comparisons for all treatment types was 30. This method was preferred over conventional MOS (mean opinion score)-based listening tests, since the scores can have a large variance due to the lack of trained listeners.

5. Results and discussion

Tables 1 and 2 list the average PESQ and segmental SNR of all enhancement methods, respectively. We can see that the Kalman oracle method achieves the highest objective scores since its AR model is estimated from the clean speech. The proposed Kalman filtering algorithm (Kalman proposed) has mostly improved PESQ scores over the Kalman iterative method, especially at the low input SNRs, while it appears to be slightly better than the MMSE-STSA method. However, in terms of segmental SNR (shown in Table 2), the Kalman proposed method can be seen to be significantly outperforming MMSE-STSA and is even competitive with the Kalman oracle method. Recent findings in the speech enhancement literature [11] have found segmental SNR to be more consistent with subjective preference scoring than PESQ. We have similarly found the segmental SNR improvements associated with the proposed Kalman filtering algorithm to be consistent with improved subjective quality in the listening tests that are described later.

In Figure 4, spectrograms are shown of the clean and noisy speech, as well as the output from the four enhancement methods. The utterance was “Clams are small, round, soft, and tasty”, where the input SNR is at 10 dB. We can see that the Kalman oracle method in Figure 4(c) exhibited the best enhancement performance, which was consistent in terms of its objective measures (PESQ and segmental SNR). However, when using the AR parameter estimates from the noisy speech, residual noise started appearing in the Kalman iterative method, as shown in Figure 4(d). The proposed method [Figure 4(e)] exhibited a comparable level of residual noise to the oracle method. The MMSE-STSA method [Figure 4(f)] appeared to suffer from a metallic residual noise in all frequency bands, hence it exhibited a lower PESQ and segmental SNR.

Figure 5 shows the mean preference scores from the subjective listening tests as well as error bars that indicate 95% confidence intervals. It can be seen that apart from the clean speech, the Kalman oracle method was the most preferred method by the listeners. The proposed method was the next preferred followed by the MMSE-STSA and Kalman iterative method. As mentioned previously, these subjective preference scores are correlated with the segmental SNR results of Table 2.

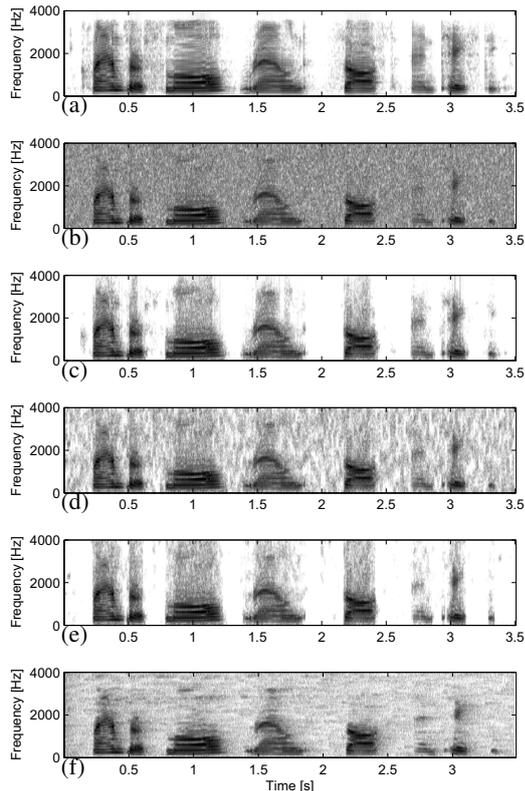


Figure 4: Spectrograms of digital speech (“Clams are small, round, soft, and tasty”): (a) with no noise; (b) corrupted by WGN at 10 dB SNR; (c) enhanced by Kalman oracle; (d) enhanced by Kalman iterative; (e) enhanced by Kalman proposed; and (f) enhanced by MMSE-STSA.

6. Conclusion

In this paper, we have presented an iterative Kalman filtering algorithm that utilises robustness and sensitivity metrics jointly to dynamically tune the Kalman filter gain to overcome poor AR parameter estimates. Both of these metrics are computed in real-time and incorporated into an iterative Kalman filtering framework. Experimental results (PESQ and segmental SNR) showed the proposed method to be competitive with the oracle-case Kalman filter and was better than the MMSE-STSA algorithm. Subjective blind listening tests also corroborated the objective findings, where the listeners preferred the enhanced speech from the proposed method over those produced by the MMSE-STSA algorithm and conventional iterative Kalman filter.

Table 1: Average PESQ results over 30 sentences from the NOIZEUS database, which compare the different speech enhancement methods with the proposed method for speech corrupted by white Gaussian noise.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	1.57	1.83	2.13	2.47
Kalman oracle	2.50	2.79	3.08	3.38
Kalman iterative	1.92	2.29	2.63	2.98
Kalman proposed	2.08	2.39	2.67	2.97
MMSE-STSA	1.96	2.33	2.64	2.94

Table 2: Average segmental SNR (in dB) results over 30 sentences from the NOIZEUS database, which compare the different speech enhancement methods with the proposed method for speech corrupted by white Gaussian noise.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	-8.31	-3.31	1.69	6.69
Kalman oracle	4.61	6.81	9.41	12.39
Kalman iterative	-0.48	3.48	7.32	11.14
Kalman proposed	3.32	5.85	8.68	11.78
MMSE-STSA	-0.32	3.15	6.40	9.40

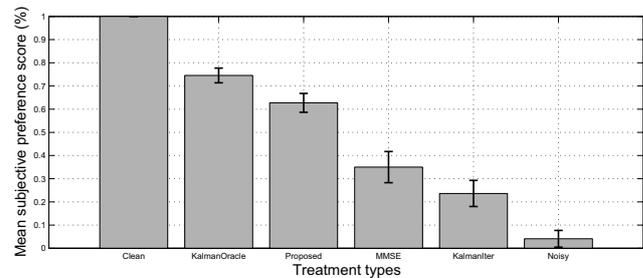


Figure 5: Mean subjective preference scores with 95% confidence intervals for all treatment types.

7. References

- [1] P. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. CRC Press LLC, 2007.
- [2] K. K. Paliwal and A. Basu, “A speech enhancement method based on Kalman filtering,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 12, Apr. 1987, pp. 177–180.
- [3] J. D. Gibson, B. Koo, and S. D. Gray, “Filtering of colored noise for speech enhancement and coding,” *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.
- [4] P. Sorqvist, P. Handel, and B. Ottersten, “Kalman filtering for low distortion speech enhancement in mobile communication,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1997, pp. 1219–1222.
- [5] S. So and K. K. Paliwal, “Suppressing the influence of additive noise on the Kalman filter gain for low residual noise speech enhancement,” *Speech Commun.*, vol. 53, no. 3, pp. 355–378, Mar. 2011.
- [6] —, “Modulation-domain Kalman filtering for single-channel speech enhancement,” *Speech Commun.*, vol. 53, no. 6, pp. 818–829, Jul. 2011.
- [7] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. New Jersey: John Wiley, 1996.
- [8] M. Saha, R. Ghosh, and B. Goswami, “Robustness and sensitivity metrics for tuning the extended Kalman filter,” *IEEE Trans. Instrum. Meas.*, vol. 63, no. 4, pp. 964–971, Apr. 2014.
- [9] S. So, A. E. W. George, R. Ghosh, and K. K. Paliwal, “A non-iterative Kalman filtering algorithm with dynamic gain adjustment for single-channel speech enhancement,” *International Journal of Signal Processing Systems*, vol. 4, no. 1, pp. 263–268, Aug. 2016.
- [10] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 1109–1121, Dec. 1984.
- [11] B. Schwerin and K. K. Paliwal, “Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement,” *Speech Commun.*, vol. 58, pp. 49–68, Mar. 2014.