GRIFFITH UNIVERSITY

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

Author

George, Aidan, So, Stephen, Ghosh, Ratna, Paliwal, Kuldip

Published

2016

Conference Title

Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology

Version

Version of Record (VoR)

Downloaded from

http://hdl.handle.net/10072/100784

Link to published version

https://assta.org/sst-2016-proceedings/

Griffith Research Online

https://research-repository.griffith.edu.au

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 filter[1].

2. Kalman filter-based speech enhancement

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

$$y(n) = x(n) + v(n) \tag{1}$$

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 pth order autoregressive (AR) model is used to represent speech production, whose parameters $\{a_k; k = 1, 2, \ldots, p\}$ 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 $\hat{\boldsymbol{x}}(n|n)$ at time n, when given the noisy speech measurement $y(n)$ and the *a priori* state vector estimate $\hat{\boldsymbol{x}}(n|n-1)$. For a detailed description of each variable, the reader is referred to [5, 7]:

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

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

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

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{A}\hat{\boldsymbol{x}}(n-1|n-1) \tag{5}$$

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

In conventional Kalman filtering, the enhanced speech signal at the present time n is given by[2]:

$$\hat{x}(n|n) = \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n) \tag{7}$$

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

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

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

where $\hat{x}(n|n-1)$ and $K(n) = \boldsymbol{c}^T\boldsymbol{K}(n)$ 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

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

[2] We 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]:

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

where $\alpha^2(n) = \boldsymbol{c}^T \boldsymbol{A}\boldsymbol{P}(n-1|n-1)\boldsymbol{A}^T \boldsymbol{c}$ 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 assumed[3] 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:

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

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

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:

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

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

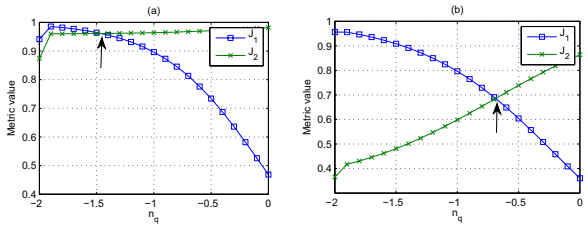

Figure 1: Plot of average Kalman filter sensitivity and robustness metrics (J_1 and J_2) over a frame for varying speech excitation variance $\hat{\sigma}_w^2 = (10^{n_q})\sigma_w^2$ (compromise point where $J_1 = J_2$ shown by the arrow) in a: (a) speech-absent region (only noise); and (b) voiced-speech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

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

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:

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

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:

$$K'(n) = [1 - J_2(n)]K(n) \tag{16}$$

One of the problems with this method was that there was oversuppression 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,

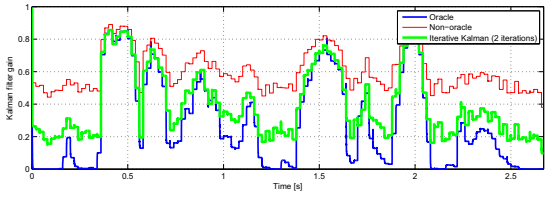

Figure 2: Plot of the scalar Kalman filter gain for oracle and non-oracle modes compared with that of the conventional iterative Kalman filter (with two iterations) [3]. Utterance was sp10 ("The sky that morning was clear and bright blue") corrupted with WGN at 5 dB SNR.

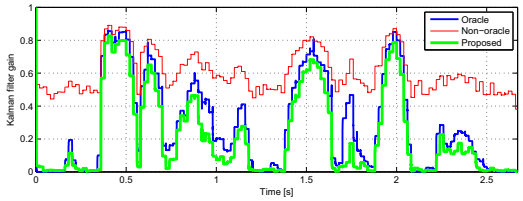

Figure 3: Plot of the scalar Kalman filter gain for oracle and non-oracle modes compared with that of the proposed iterative Kalman filter (with two iterations). Utterance was sp10 ("The sky that morning was clear and bright blue") corrupted with WGN at 5 dB SNR.

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 $\hat{\sigma}^2_{wc}$ 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.

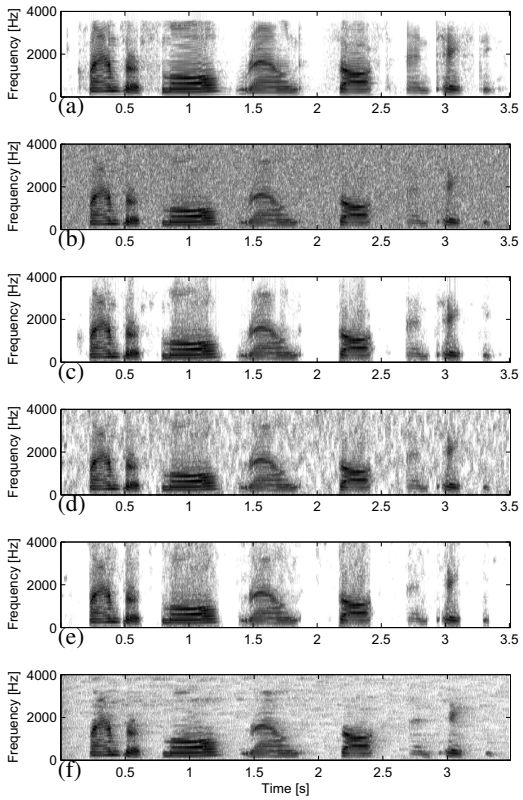

Figure 4: Spectrograms of digital speech ("Clams are small, round, soft, and tasty"): (a) with no noise; (b) corrupted by WGN at 10 dB SNR; (c) enhanced by Kalman oracle; (d) enhanced by Kalman iterative; (e) enhanced by Kalman proposed; and (f) enhanced by MMSE-STSA.

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

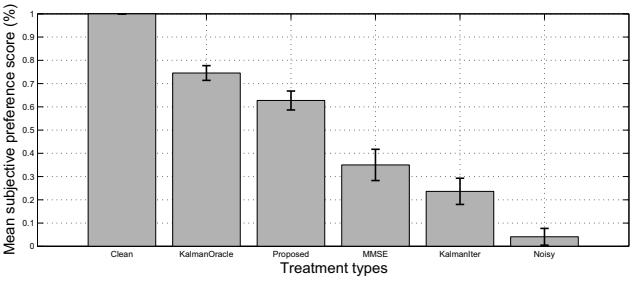

Figure 5: Mean subjective preference scores with 95% confidence intervals for all treatment types.

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.

A comparison of estimation methods in the discrete cosine transform modulation domain for speech enhancement

Aidan E.W. George, Christine Pickersgill, Belinda Schwerin, Stephen So

Griffith School of Engineering, Gold Coast campus,
Griffith University, QLD, 4222 Australia.

{aidan.george, christine.pickersgill}@griffithuni.edu.au,
{b.schwerin, s.so}@griffith.edu.au

Abstract

In this paper, we present a new speech enhancement method that processes noise-corrupted speech in the discrete cosine transform (DCT) modulation domain. In contrast to the Fourier transform, the DCT produces a real-valued signal. Therefore, modulation-based processing in the DCT domain may allow both acoustic Fourier magnitude and phase information to be jointly estimated. Based on segmental SNR and the results of blind subjective listening tests on various coloured noises, the application of the subspace method in the DCT modulation domain processing was found to outperform all other methods evaluated, including the LogMMSE method.

Index Terms: speech enhancement, discrete cosine transform, modulation domain, subspace method

1. Introduction

Environmental noise is an ever-present problem in many speech processing applications, such as speech coding and automatic speech recognition (or ASR), since its presence in the captured speech signal adversely impacts on performance. For instance, in speech coding, the quality and intelligibility of the decoded speech degrades as the signal-to-noise ratio (or SNR) of the input signal decreases. In ASR tasks, low input SNR leads to a deterioration of the recognition accuracy.

One technique for addressing the problem of noise is to use speech enhancement. The role of speech enhancement is to reduce the level of undesirable noise in a speech signal in order to improve the quality and intelligibility. Several speech enhancement algorithms have been reported in the literature, each of which process speech in different domains. Early enhancement methods, such as spectral subtraction [1], MMSE-STSA [2], and LogMMSE [3], estimated the clean speech in the acoustic magnitude domain. In subspace enhancement methods, a linear estimator is applied in the signal and noise subspace domains. Recently, more successful speech enhancement methods have been reported that apply spectral subtraction [4], MMSE [5] and Kalman filters [6] in the modulation magnitude domain.

Typically, in modulation domain methods, an analysis-modification-synthesis (or AMS) procedure is first performed and estimation is applied on the temporal trajectories of Fourier transform magnitudes at each acoustic frequency [4]. The enhanced acoustic Fourier magnitude information is then combined with the noisy acoustic Fourier phase and synthesised using a windowed overlap-adding procedure to obtain the enhanced speech frame. A common characteristic of the acoustic and modulation domain methods is the use of noisy and unprocessed Fourier phase information in the synthesis stage. A recent study [7] showed that at low instantaneous spectral SNR that are less than 7 dB, noise in the phase information has a notable effect on the quality and intelligiblity of the enhanced speech output. Therefore, it is advantageous to process both Fourier magnitude and phase information.

One method of incorporating phase information into the enhancement procedure is to process real and imaginary parts (RI) of the complex Fourier transform [8, 9], which is also referred to as RI modulation. This is based on the concept that the real and imaginary parts contribute to both the magnitude and phase components. In these methods, an estimator is *applied independently* on the real and imaginary modulation signals over time. However, one may speculate that *joint processing* of RI modulation signals is advantageous in better estimating the clean phase information.

In this paper, we report on a new speech enhancement approach that estimates modulation magnitude and phase information in the discrete cosine transform (DCT) domain. It is well known that the DCT is a unitary transform that outputs a real-valued signal, in contrast to the discrete Fourier transform, which outputs a complex-valued signal. This suggests that the Fourier magnitude and phase information is embedded within the single real-valued signal. Therefore, a speech enhancement approach that is based on processing in the DCT modulation domain can be viewed as a *joint estimator* of magnitude and phase information. In order to determine an effective estimation method in the DCT modulation domain, our study evaluated three methods: spectral subtraction [1], LogMMSE [3] and the subspace approach [10]. Speech enhancement experiments on the NOIZEUS speech corpus [11] were performed to evaluate the different methods in the DCT modulation domain as well as compare them against the conventional subspace and acoustic magnitude domain methods, such as spectral subtraction and LogMMSE[1]. Objective scores and blind subjective listening tests suggested that the subspace approach was the more effective method for estimation in the DCT modulation domain. Furthermore, they indicated that the proposed DCT-SSp enhancement method produced better quality enhanced speech than the spectral subtraction, LogMMSE and the conventional subspace method.

2. Method

In this work we propose a modulation-domain speech enhancement method which makes use of the discrete cosine transform (DCT) in place of the short-time Fourier Transform (STFT). The DCT has the advantage in that it provides a frequency representation of a signal where the magnitude and phase information are jointly represented by a single real-valued spectrum. The DCT has been particularly popular in speech and image coding due to its ability to concentrate signal energy in a few coefficients, but has also been applied to speech enhancement. For example, in [12] the MMSE method was used to estimate the DCT coefficients on a frame-by-frame basis. In contrast, we have investigated the use of DCT in a modulation-domain based framework where DCT coefficient time trajectories are processed in order to suppress noise. Further, we demonstrate that the subspace enhancement approach is well suited to enhancing speech in this domain.

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

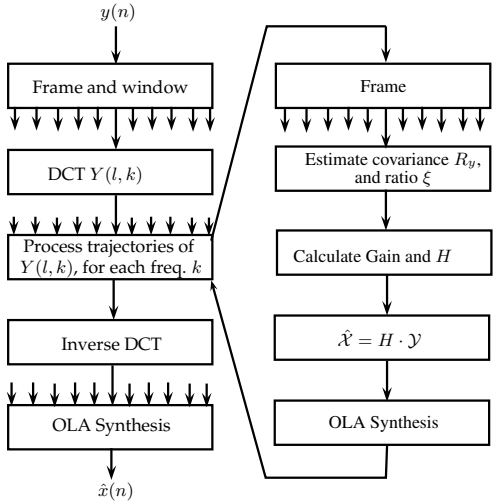

Figure 1: *DCT-SSp method for enhancing speech using a DCT-based modulation framework (left) and suppression of noise in the DCT spectral time trajectories using a subspace estimation method (right).*

2.1. DCT-based modulation domain framework

Consider a speech signal $x(n)$ corrupted by additive noise $d(n)$ to give the noisy signal $y(n) = x(n) + d(n)$. Assuming speech to be quasi-stationary, the noisy speech signal can be analysed framewise by applying the DCT to each windowed frame of the signal, to give the noisy DCT spectrum

$$Y(l,k) = u(k) \sum_{n=0}^{N-1} y(n+lZ) v_a(n) \cos\left[\frac{\pi(2n-1)k}{2N}\right] \qquad k = 0, 1, ..., N-1, \quad (1)$$

where

$$u(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 0, \\ \sqrt{\frac{2}{N}}, & 1 \le k \le N-1. \end{cases} \quad (2)$$

Here l refers to the frame index, k is the frequency index, Z is the frame shift, N is the frame length, and $v_a(n)$ is the analysis window.

The time trajectories for each frequency component of the DCT spectra $Y(l,k)$ form a modulation signal which can then be processed framewise to suppress noise. The enhanced speech signal $\hat{x}(n)$ can then be reconstructed from the modified spectra $\hat{X}(l,k)$ using an inverse DCT, followed by overlap-add (OLA) synthesis. That is,

$$\hat{x}(n) = \sum_{l} v_s(n - lZ) \sum_{k=0}^{N-1} u(k) \hat{X}(l,k) \cos\left[\frac{\pi(2n-1)k}{2N}\right], \qquad k = 0, 1, ..., N-1, \quad (3)$$

where $v_s(n)$ is the synthesis window.

2.2. Subspace enhancement of the modulation signal

To enhance speech within the DCT-based modulation framework described in Section 2.1, each modulation signal is processed within a separate framework where the signal is framed and the generalized subspace method for enhancing speech corrupted by coloured noise [10] is used, as represented in Figure 1.[2] Processing of the modulation signal begins with overlapped framing, then the following procedure is used to modify each frame of the signal.

The covariance matrix R_y of each noisy frame is estimated using the multi-window method [13] and sine tapers [14]. For this approach, the number of tapers used has a considerable effect on the resulting quality of enhanced stimuli. Using only a small number of tapers results in increased variance in the estimate, while a higher number of tapers results in reduced variance but reduced resolution (or larger bias). Thus, the number of tapers used in estimating the covariance matrix provides a trade-off between different distortion types.

The ratio of the clean to noise covariance ξ is then estimated from the noisy and noise covariance matrices as

$$\xi = R_n^{-1} R_x = R_n^{-1} R_y - I, \quad (4)$$

where the noise covariance matrix R_n is initially estimated from the leading noise-only region of the utterance (using the multi-window method), and updated in non-speech frames (as determined by a simple SNR-based VAD calculated from the first coefficient of the noisy and noise covariance matrices).

Eigen-decomposition is then used to find the eigenvalues Λ and eigenvectors V of ξ,

$$\xi V = V \Lambda. \quad (5)$$

The dimension of the speech signal subspace M is then given by the number of non-zero positive eigenvalues of ξ.

The gain matrix can then be calculated as

$$g_{kk} = \frac{\lambda(k)}{\lambda(k) + \mu}, \qquad k = 1, 2, ..., M$$
$$G = \text{diag}\{g_{11}, g_{22}, ..., g_{MM}\} \quad (6)$$

where $\lambda(k)$ are the positive non-zero eigenvalues of ξ (in decreasing order). In Eq. (6), μ provides a trade-off between speech distortions and residual noise, and is selected based on SNR as

$$\mu = \begin{cases} \mu_o - \frac{\text{SNR}_{dB}}{slope}, & -5 < \text{SNR}_{dB} < 20 \\ 1, & \text{SNR}_{dB} \ge 20 \\ 5, & \text{SNR}_{dB} \le -5. \end{cases} \quad (7)$$

where SNR_{dB} is estimated from the eigenvalues corresponding to the speech signal subspace (and therefore estimating the eigenvalues of R_x), and μ_o and $slope$ are empirically determined values. Here, $\mu_o = 4.2$ and $slope = 0.16$, as suggested in [10] were used.

The estimate of the enhanced frame is then given by

$$\hat{\mathcal{X}}(\ell, m, k) = H \cdot \mathcal{Y}(\ell, m, k) v_m(k), \quad (8)$$

where

$$H = V^{-T} \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} V^T, \quad (9)$$

and where ℓ is the modulation signal frame index, m is the frame value index, and $v_m(k)$ is the window function. Finally, OLA synthesis is used to reconstruct the modified modulation signal $\hat{X}(l,k)$.

3. Experiments

A number of experiments were conducted to evaluate the quality of stimuli processed using the proposed DCT-SSp method. Both blind subjective tests and objective evaluations were performed. Degraded stimuli used in experiments were stimuli from the Noizeus speech corpus [11] corrupted by F16, Babble and Car noises, at input noise levels ranging from 0 to 15 dB. For subjective evaluations, blind AB listening tests were conducted, in which 16 English-speaking listeners compared the quality of stimuli corrupted by F16 noise at 5 dB and processed using different treatment types, and selected the stimuli they preferred. These listening tests made use of two sentences, one from a male and one from a female speaker, and involved listening to 84 stimuli pairs per experiment. Objective experiments compared stimuli processed using each treatment type to the original clean stimuli, for a range of input SNR values. Mean

[2] For a more detailed account of the generalized subspace approach for speech corrupted by colour noise, the reader is referred to [10].

Table 1: *Parameters used in implementing the proposed DCT-SSp enhancement method.*

Parameter	DCT	Subspace
Frame duration	20 ms	16 ms
Sub-frame duration	-	2 ms
Frame update	0.625 ms	8 ms
Analysis window	Hamming	-
Synthesis window	modified Hann	Hamming

Segmental SNRs for each treatment type and input SNR were then calculated across the 30 utterances of the Noizeus corpus.

An important parameter in the DCT-SSp method is the number of tapers used in the estimation of the covariance matrix, since the variance-bias tradeoff has a significant effect on the quality. A small number of tapers leads to good speech quality but a high level of residual noise, while a large number of tapers results in a more distorted speech along with reduced levels of residual noise. Blind listening tests performed by a subset of listeners found that 32 and 128 tapers were the most preferred, with some listeners preferring lower residual noise levels and others preferring less speech distortion. We have included both in the formal listening tests, where stimuli generated using 32 tapers are denoted DCT-SSp32 and those generated using 128 tapers are denoted DCT-SSp128. Other parameters used in the construction of DCT-SSp stimuli are shown in Table 1.

In the first round of listening tests, the suitability of the subspace method in the DCT modulation domain was examined. Treatment types included in the subjective tests were the proposed subspace method (DCT-SSp), spectral subtraction (DCT-SpSub), and LogMMSE (DCT-LogMMSE). The DCT-SpSub and DCT-LogMMSE methods utilised the DCT framework described in Section 2.1. Each trajectory was then processed using either the spectral subtraction method of Boll [1] or the LogMMSE magnitude estimator of Ephraim and Malah [3].

In the second set of listening tests, the proposed DCT-SubSpace method was compared with popular acoustic domain methods including spectral subtraction, LogMMSE, and subspace. Acoustic domain enhancement methods were implemented using publicly available reference implementations [11].

4. Results and discussion

Mean subjective preference scores for the first listening test for utterances sp10 (male speaker) and sp15 (female speaker) are shown in Figure 2. Stimuli included in subjective evaluations were clean, noisy, DCT modulation domain spectral subtraction (DCT-SpSub), LogMMSE (DCT-LogMMSE), and Subspace using 32 tapers (DCT-SSp32) and 128 tapers (DCT-SSp128). Scores shown in Figure 2 indicate that both DCT Subspace treatment types were preferred over DCT LogMMSE and DCT spectral subtraction methods.

Separate one-way ANOVA tests were performed on the male and female spoken stimuli sets to determine the statistical significance of the first round of subjective results. The null hypothesis ("all means are equal") was found to have been rejected (for Sp10, F(6, 105) = 130.26, $p < 0.05$; and for Sp15, F(6, 105) = 110.76, $p < 0.05$). Post hoc analysis using the Tukey's Honestly Significant Difference (HSD) tests performed on each stimuli set found most treatment types to be significantly different, with the exception of the comparison between DCT-SSp32 and DCT-SSp128.

A second listening test was performed using clean, noisy, and stimuli processed with acoustic spectral subtraction (SpecSub), acoustic Log-MMSE (LogMMSE), acoustic Subspace using 16 tapers(Ac-SSp), DCT-SSp32, and DCT-SSp128. Mean subjective preference scores given in Figure 3 show that the two DCT Subspace methods are preferred among listeners. Though

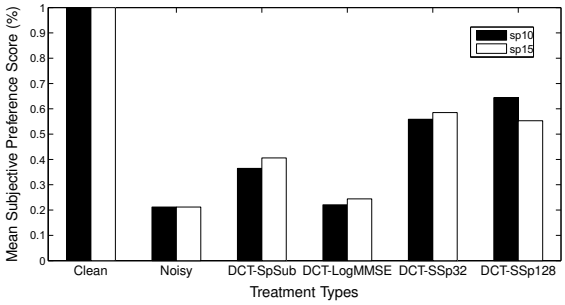

Figure 2: *Average subjective test results featuring clean, noisy (5 dB F16 noise), and stimuli processed with DCT-SpSub, DCT-LogMMSE, DCT-SSp32, and DCT-SSp128.*

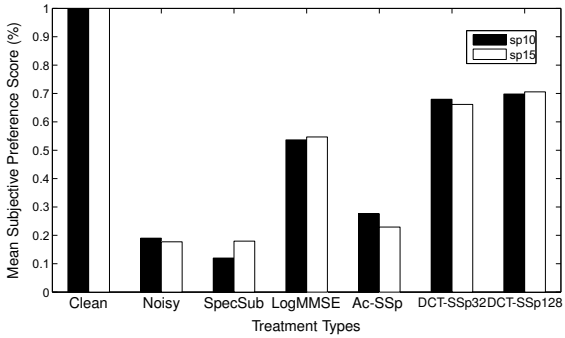

Figure 3: *Average subjective test results featuring clean, noisy (5 dB F16 noise), and stimuli processed with SpecSub, LogMMSE, Ac-SSp, DCT-SSp32, and DCT-SSp128.*

LogMMSE also shows a high percentage of preference, the post hoc HSD analysis found that it was significantly different to both the DCT-SSp treatment types. Identical ANOVA and post hoc HSD tests performed on both the male and female spoken stimuli sets. The null hypothesis was found to have been rejected (for Sp10, F(5, 96) = 68.0, $p < 0.05$; and for Sp15, F(5, 96) = 55.67, $p < 0.05$) and post hoc analysis identified significant differences between all treatment types, though not between DCT-SSp32 and DCT-SSp128.

Spectrograms (for utterance sp10) of stimuli used in all subjective quality tests are shown in Figure 4. Clean and noisy (5 dB added F16 noise) are shown in Figures 4(a) and (b) respectively. We can see that the acoustic domain methods (SpecSub and LogMMSE) in Figures 4(c) and (d) suffer from medium levels of residual noise that are characteristic of the respective methods. The conventional Ac-SSp method (Figure 4(e)) exhibits better noise suppression, though the residual noise is rather non-stationary and the speech has suffered from some spectral smoothing, giving it a 'breathy' nature. For the DCT modulation methods, it can be seen that the proposed DCT-SSp32 and DCT-SSp128 methods (Figures 4(h) and (i)) exhibit much better noise suppression than DCT-LogMMSE and DCT-SpSub. The speech is clearer and less bottled for DCT-SSp32, but also contains a residual tone which is not audible in DCT-SSp128. While spectrograms for utterance sp10 are shown here, observations of spectrograms for other utterances from the Noizeus corpus corrupted with 5 dB F16 noise also show similar properties.

Objective quality evaluations were performed for all treatment types at various input SNR levels and for different types of coloured noise. Segmental SNR scores for three noise types (F16, Babble, and Car) are given in Table 2. In most cases the two proposed methods have scored higher than the other treat-

Table 2: *Segmental SNR scores for all treatment types at various input SNR levels and for three types of coloured noise.*

Method	F16				BABBLE				CAR			
	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
Noisy	-4.329	-1.445	1.763	5.245	-4.098	-1.153	1.972	5.563	-4.253	-1.423	1.807	5.355
SpecSub	-0.149	2.801	6.000	9.307	-0.739	2.132	5.109	8.442	0.178	3.004	6.010	9.464
LogMMSE	0.410	3.051	5.634	8.153	-1.234	1.500	4.200	6.917	0.474	3.001	5.580	8.231
Ac-SSp	0.774	3.237	5.495	7.675	-0.363	2.155	4.408	6.675	1.044	3.302	5.372	7.666
DCT-SpSub	-1.932	-0.870	-0.019	0.573	-3.137	-1.438	-0.695	0.261	-1.711	-0.673	-0.014	0.707
DCT-LogMMSE	-5.199	-5.069	-4.956	-4.828	-5.455	-5.186	-5.080	-4.928	-5.152	-5.028	-4.953	-4.813
DCT-SSp32	1.723	4.283	6.794	9.109	-0.174	2.393	5.312	7.814	1.603	4.066	6.497	8.987
DCT-SSp128	2.428	4.692	7.053	9.214	0.302	2.713	5.584	8.011	2.293	4.437	6.667	9.040

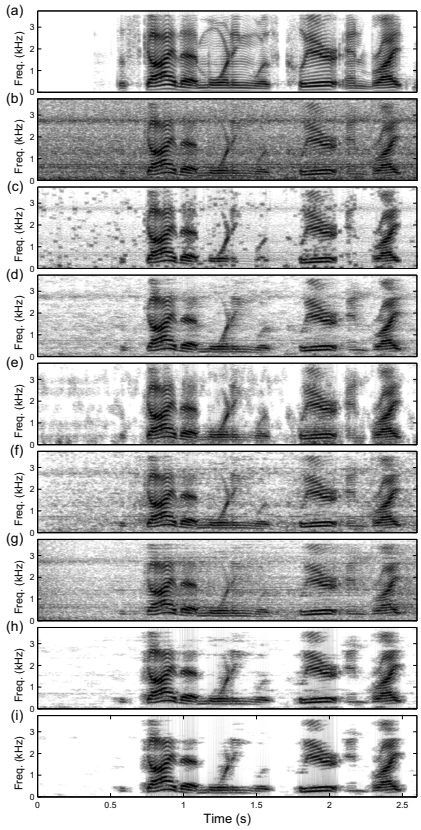

Figure 4: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech; (b) speech degraded by 5 dB F16 noise; (c) Spec-Sub; (d) LogMMSE; (e) Ac-SSp; (f) DCT-SpSub; (g) DCT-LogMMSE; (h) DCT-SSp32; (i) DCT-SSp128.*

ment types, often by a considerably large margin. Scores for F16 noise accurately reflect the subjective test results shown in Figure 2 and Figure 3.

5. Conclusion

In this paper, we have investigated a new speech enhancement method that processes noise-corrupted speech in the discrete cosine transform modulation domain. Three enhancement methods (spectral subtraction, LogMMSE, and subspace) that processed DCT spectral coefficients temporally were evaluated using blind subjective listening tests and objective quality measures. It was found that the best enhancement performance was achieved when using the subspace method in the DCT modulation domain.

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, 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.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec 1984.

[3] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr 1985.

[4] K. Paliwal, B. Schwerin, and K. Wójcicki, "Modulation domain spectral subtraction for speech enhancement," in *Proc. ISCA Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brighton, U.K., Sep 2009, pp. 1327–1330.

[5] ——, "Speech enhancement using minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, no. 2, pp. 282–305, Feb 2012.

[6] S. So and K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Communication*, vol. 53, no. 6, pp. 818–829, July 2011.

[7] R. Chappel, B. Schwerin, and K. Paliwal, "Phase distortion resulting in a just noticeable difference in the perceived quality of speech," *Speech Communication*, vol. 81, pp. 138–147, July 2016.

[8] Y. Zhang and Y. Zhao, "Spectral subtraction on real and imaginary modulation spectra," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, May 2011, pp. 4744–4747.

[9] B. Schwerin and K. Paliwal, "Speech enhancement using STFT real and imaginary parts of modulation signals," in *Proc. of SST*, Sydney, Australia, Dec 2012, pp. 25–28.

[10] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, July 2003.

[11] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: Taylor and Francis, 2007.

[12] I. Soon, S. Koh, and C. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Communication*, vol. 24, pp. 249–257, 1998.

[13] L. McWorter and L. Scharf, "Multiwindow estimators of correlation," *IEEE Trans. Signal Process.*, vol. 46, no. 2, pp. 440–448, Feb 1998.

[14] K. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 188–195, Jan 1995.