Single-channel speech enhancement using Kalman filtering in the modulation domain

Stephen So, Kamil K. Wójcicki, Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering,
Griffith University, Brisbane, QLD, Australia, 4111
{s.so, k.wojcicki, k.paliwal}@griffith.edu.au

Abstract

In this paper, we propose the modulation-domain Kalman filter (MDKF) for speech enhancement. In contrast to previous modulation domain-enhancement methods based on bandpass filtering, the MDKF is an adaptive and linear MMSE estimator that uses models of the temporal changes of the magnitude spectrum for both speech and noise. Also, because the Kalman filter is a joint magnitude and phase spectrum estimator, under non-stationarity assumptions, it is highly suited for modulation-domain processing, as modulation phase tends to contain more speech information than acoustic phase. Experimental results from the NOIZEUS corpus show the ideal MDKF (with clean speech parameters) to outperform all the acoustic and time-domain enhancement methods that were evaluated, including the conventional time-domain Kalman filter with clean speech parameters. A practical MDKF that uses the MMSE-STSA method to enhance noisy speech in the acoustic domain prior to LPC analysis was also evaluated and showed promising results.

Index Terms: speech enhancement, Kalman filtering, modulation domain

1. Introduction

In the problem of speech enhancement, where a speech signal is corrupted by noise, we are primarily interested in suppressing the noise so that the quality and intelligibility of speech are improved. Speech enhancement is useful in many applications where corruption by noise is undesirable and unavoidable. The Kalman filter was first introduced for speech enhancement in [1], where significant noise reduction was reported when linear prediction coefficients (LPCs) estimated from clean speech were provided. In practice though, poor parameter estimates from noisy speech result in degraded enhancement performance. Iterative Kalman filters (such as [2]) have been shown to alleviate the effects of poor parameter estimates in the Kalman filter, but in some of these methods, convergence is not guaranteed and the enhanced output suffers from musical noise and speech distortion.

The Kalman filter is an unbiased, time-domain, linear minimum mean squared error (MMSE) estimator, where the enhanced speech is recursively estimated on a sample-by-sample basis. Hence, the Kalman filter can be viewed as a joint estimator for both the magnitude and phase spectrum of speech, under non-stationarity assumptions [3]. This is in contrast to the short-time Fourier transform (STFT)-based enhancement methods, such as spectral subtraction, Wiener filtering, and MMSE estimation [4], where the *noisy* phase spectrum is combined with the estimated clean magnitude spectrum to produce the enhanced speech frame. However, it has been reported that for spectral SNRs greater than approximately 8 dB, the use of unprocessed noisy phase spectrum does not lead to perceptible distortion [4].

There has been recent interest in using the modulation domain as an alternative to the acoustic domain for speech enhancement, where we define the *acoustic frequencies* as the STFT of a signal and the *modulation domain* as the temporal trajectory of the magnitude spectrum at all acoustic frequencies [5]. This is because there is growing psychoacoustic and physiological evidence to support the significance of the modulation domain for speech analysis and processing [6]. Drullman et al. [7, 8] investigated the importance of modulation frequencies for intelligibility by applying low-pass and high-pass filters to the temporal envelopes of acoustic frequency subbands. They showed frequencies between 4 and 16 Hz to be important for intelligibility, with the region around 4–5 Hz being the most significant. In a similar study, Arai et al. [9] showed that applying passband filters between 1 and 16 Hz does not impair speech intelligibility. While the envelope of the acoustic magnitude spectrum represents the shape of the vocal tract, the modulation spectrum represents how the vocal tract changes as a function of time. It is these temporal changes that convey most of the linguistic information (or intelligibility) of speech.

Hermansky et al. [10] proposed to bandpass filter the time trajectories of cubic-root compressed short-time power spectrum for enhancement of speech corrupted by additive noise. Similar bandpass filtering was applied to the time trajectories of the short-time power spectrum for speech enhancement in [11, 12]. These bandpass filtering methods have several limitations: (1) the filters are fixed in nature, while the properties of speech and noise change over time; (2) the properties of the noise are not exploited in the design of the filters; and (3) noise contained in the filter passband (the speech modulation regions) is preserved. These limitations were addressed recently in [13], whereby the spectral subtraction algorithm was used to process the modulation spectrum on a frame-by-frame basis. The results from this study demonstrated the effectiveness of speech enhancement in the modulation domain, where 'musical noise' that is typically associated with spectral subtraction in the acoustic domain, could be suppressed in the enhanced speech by increasing the modulation frame duration.

In this paper, we propose the use of Kalman filtering for estimating the temporal trajectories of the magnitude spectrum along each acoustic frequency. We believe the ability of the Kalman filter to estimate both the magnitude and phase spectrum, under non-stationarity assumptions [3] makes it preferable over STFT-based enhancement methods, because phase information has been shown to play a more important role in the modulation domain than in the acoustic domain [10]. Using objective tests on the NOIZEUS speech corpus [4], we show that in the ideal case where accurate model parameters are available, the modulation domain Kalman filter (MDKF) outperforms all acoustic and time-domain speech enhancement methods that were evaluated (including the time-domain Kalman filter (TDKF)), in the ideal case where accurate model parameters are available. We also present some results of a practical MDKF that uses the MMSE-STSA algorithm in the acoustic domain as a preprocessor for LPC estimation.

2. Modulation domain Kalman filtering for speech enhancement

2.1. Acoustic analysis-modification-synthesis framework

The analysis-modification-synthesis (AMS) framework consists of three stages: (1) the analysis stage, where the input speech is processed using STFT analysis; (2) the modification stage, where the noisy spectrum undergoes some kind of modification; and (3) the synthesis stage, where the inverse STFT is followed by the overlap-add synthesis to reconstruct the output signal.

Let us consider an additive noise model:

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

where $y(n)$, $x(n)$ and $v(n)$ denote zero-mean signals of noisy speech, clean speech and noise, respectively. Since speech can be assumed to be quasi-stationary, it is analysed framewise using short-time Fourier analysis. The STFT of the corrupted speech signal $y(n)$ is given by:

$$Y(n,k) = \sum_{l=-\infty}^{\infty} y(l)w(n-l)e^{-j\frac{2\pi kl}{N}} \tag{2}$$

where k refers to the index of the discrete acoustic frequency, N is the acoustic frame duration (in samples) and $w(n)$ is an acoustic analysis window function. In speech processing, the Hamming window with 20–40 ms duration is typically employed. Using STFT analysis, we can represent Eq. (2) as:

$$Y(n,k) = X(n,k) + V(n,k) \tag{3}$$

where $Y(n,k)$, $X(n,k)$ and $V(n,k)$ are the STFTs of noisy speech, clean speech, and noise, respectively. Each of these can be expressed in terms of acoustic magnitude and acoustic phase spectrum. For instance, the STFT of the noisy speech signal can be written in polar form as:

$$Y(n,k) = |Y(n,k)|e^{j\angle Y(n,k)} \tag{4}$$

where $|Y(n,k)|$ denotes the acoustic magnitude spectrum and $\angle Y(n,k)$ denotes the acoustic phase spectrum.

Traditional AMS-based speech enhancement methods modify, or enhance, only the noisy acoustic magnitude spectrum while keeping the noisy acoustic phase spectrum unchanged. Let us denote the enhanced magnitude spectrum as $|\hat{X}(n,k)|$, then the modified acoustic spectrum is constructed by combining $|\hat{X}(n,k)|$ with the noisy phase spectrum, as follows:

$$\hat{X}(n,k) = |\hat{X}(n,k)|e^{j\angle Y(n,k)} \tag{5}$$

The enhanced speech $\hat{x}(n)$ is reconstructed by taking the inverse STFT of the modified acoustic spectrum followed by synthesis windowing and overlap-add reconstruction [14].

2.2. Kalman filtering in the modulation domain

The modulation domain views the acoustic magnitude spectrum as a series of N *modulating signals* that span across time. Each modulating signal represents the temporal evolution of each acoustic magnitude spectral component, as shown in Fig. 2. In the proposed modulation-domain Kalman filter (MDKF), each modulating signal, $|Y(n,k)|$ (where $k = 1, 2 \dots, N$) is processed using a Kalman filter (see Fig. 1).

In the modulation-domain Kalman filter, we assume an additive noise model for each modulating signal:

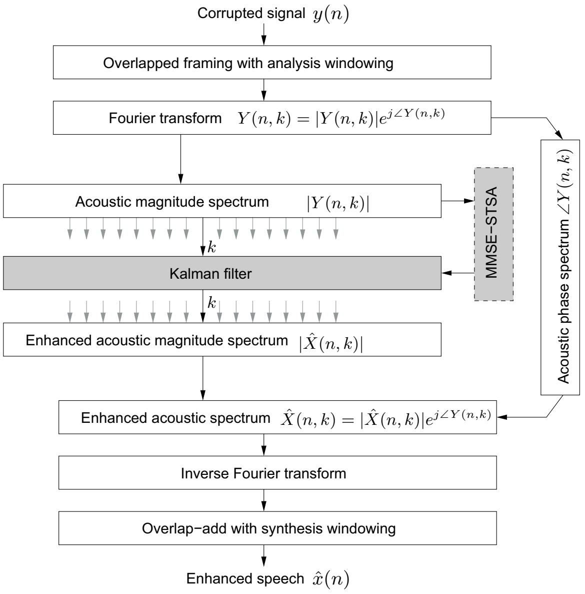
$$|Y(n,k)| = |X(n,k)| + |V(n,k)| \tag{6}$$

where $|V(n,k)|$ is the kth modulating signal of white Gaussian noise. A pth order linear predictor can be used to model the kth modulating signal of speech and together with the corrupting noise, we can write the following state space representation for $|Y(n,k)|$:

$$\boldsymbol{X}(n,k) = \boldsymbol{A}(k)\boldsymbol{X}(n-1,k) + \boldsymbol{d}W(n,k) \tag{7}$$

$$|Y(n,k)| = \boldsymbol{c}^T\boldsymbol{X}(n,k) + |V(n,k)| \tag{8}$$

where $\boldsymbol{A}(k)$ is the state transition matrix, $\boldsymbol{X}(n,k) = [|X(n,k)|, |X(n-1,k)|, \dots, |X(n-p+1,k)|]^T$ is the clean modulation state vector, $\boldsymbol{d} = [1, 0, \dots, 0]^T$ and $\boldsymbol{c} = [1, 0, \dots, 0]^T$ are the measurement vectors for the excitation noise $W(n,k)$ and observation, respectively.



Corrupted signal $y(n)$
Overlapped framing with analysis windowing
Fourier transform $Y(n,k) = |Y(n,k)|e^{j\angle Y(n,k)}$
Acoustic magnitude spectrum $|Y(n,k)|$
MMSE–STSA
Acoustic phase spectrum $\angle Y(n,k)$
k
Kalman filter
k
Enhanced acoustic magnitude spectrum $|\hat{X}(n,k)|$
Enhanced acoustic spectrum $\hat{X}(n,k) = |\hat{X}(n,k)|e^{j\angle Y(n,k)}$
Inverse Fourier transform
Overlap−add with synthesis windowing
Enhanced speech $\hat{x}(n)$

Figure 1: Schematic diagram of the proposed AMS-based modulation-domain Kalman filtering framework.

The Kalman filter recursively computes an unbiased and linear MMSE estimate $\hat{\boldsymbol{X}}(n|n,k)$ of the kth modulation state vector at time n, given the noisy modulating signal $|Y(n,k)|$, by using the following equations:

$$\boldsymbol{P}(n|n-1,k) = \boldsymbol{A}(k)\boldsymbol{P}(n-1|n-1,k)\boldsymbol{A}(k)^T + \sigma^2_{W(k)}\boldsymbol{d}\boldsymbol{d}^T \tag{9}$$

$$\boldsymbol{K}(n,k) = \boldsymbol{P}(n|n-1,k)\boldsymbol{c}\left[\sigma^2_{V(k)} + \boldsymbol{c}^T\boldsymbol{P}(n|n-1,k)\boldsymbol{c}\right]^{-1} \tag{10}$$

$$\hat{\boldsymbol{X}}(n|n-1,k) = \boldsymbol{A}(k)\hat{\boldsymbol{X}}(n-1|n-1,k) \tag{11}$$

$$\boldsymbol{P}(n|n,k) = [\boldsymbol{I} - \boldsymbol{K}(n,k)\boldsymbol{c}^T]\boldsymbol{P}(n|n-1,k) \tag{12}$$

$$\hat{\boldsymbol{X}}(n|n,k) = \hat{\boldsymbol{X}}(n|n-1,k) + \tag{13}$$

$$\boldsymbol{K}(n,k)\left[|Y(n,k)| - \boldsymbol{c}^T\hat{\boldsymbol{X}}(n|n-1,k)\right] \tag{14}$$

When applying the Kalman filter in the modulation domain, there are some time domain-based assumptions that may not necessarily be satisfied in the modulation domain:

- additive noise in the time domain may not be additive in the modulation domain (Eq. (6));
- white noise in the time domain may not be spectrally white in the modulation domain; and
- the linear predictor may not be the best dynamic model of modulating signals.

In regards to the additive noise assumption in the modulation domain, let us consider Eq. (2) in polar form:

$$|Y(n,k)|e^{j\angle Y(n,k)} = |X(n,k)|e^{j\angle X(n,k)} + |V(n,k)|e^{j\angle V(n,k)} \tag{15}$$

Using a geometric approach [4], it is easy to see that the additive noise assumption of Eq. (6) is approximately satisfied if either $\angle X(n,k) \approx \angle V(n,k)$ or $|X(n,k)| >> |V(n,k)|$. The first condition is more difficult to show since it is assumed that clean speech and noise signals are not correlated. However, the second condition is related to the instantaneous spectral SNR at acoustic frequency index k, i.e. $|X(n,k)|^2/|V(n,k)|^2$. Hence it can be inferred that the additive noise assumption in the modulation domain is roughly satisfied in high spectral SNR regions.

Fig. 3 shows the autocorrelation function of the modulating signal at an acoustic frequency for white Gaussian noise.

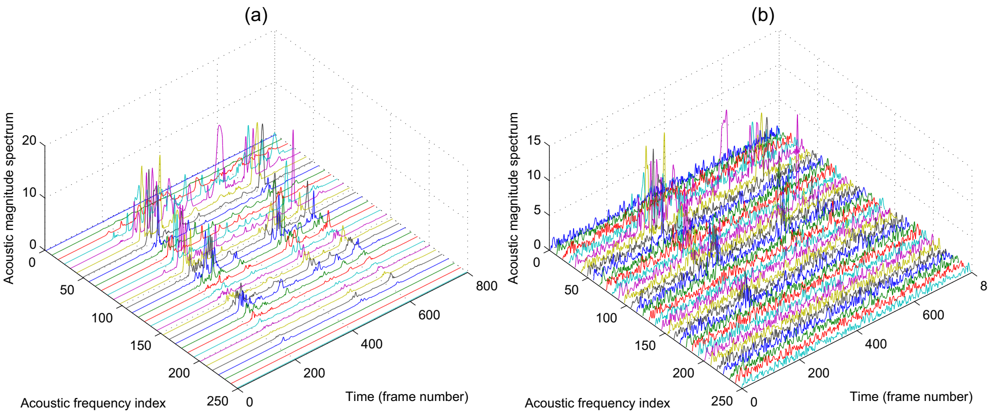


Figure 2: The modulation domain representation of speech ('The sky that morning was clear and bright blue'): (a) clean speech; (b) speech corrupted with white Gaussian noise at an SNR of 0 dB.

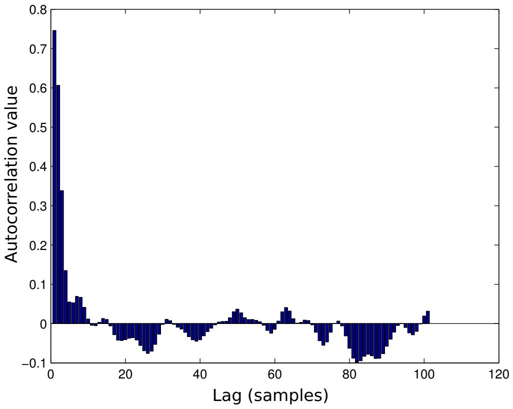


Figure 3: Plot of autocorrelation function of the modulating signal at an acoustic frequency of white Gaussian noise.

We can see that the modulating signals of white noise do contain some correlation at higher lags and hence their modulation spectrum is not white. Therefore, in order to accommodate this fact, the coloured-noise Kalman filter [2] is chosen for use in the proposed MDKF, where an extra qth linear predictor is used to model the noise and the state vectors and transition matrices are augmented to sizes of $p + q$.

Finally, in regards to the dynamic model, we have observed in our experiments that for the MDKF in the ideal case (where clean speech parameters are available), the linear predictor is sufficient at modelling the modulating signals of clean speech. However, it is well known that the presence of noise will introduce bias in the LPC estimates, which degrades the performance of the Kalman filter. In the proposed MDKF (see Fig. 1), we employ the MMSE-STSA method in the acoustic domain to enhance the speech prior to LPC estimation in the modulation domain, in order to reduce the effect of noise. We should note that other models may be more applicable for predicting the temporal evolution of these modulating signals in the presence of noise and these will be investigated in a future study.

3. Speech enhancement experiments

3.1. Experimental setup

In our experiments, we use the NOIZEUS speech corpus, which is composed of 30 phonetically balanced sentences belonging to six speakers [4]. The corpus is sampled at 8 kHz. For our objective experiments, we generate a stimuli set that has been corrupted by additive white Gaussian noise at four SNR levels (0, 5, 10 and 15 dB). The noise-only sections of all the stimuli have been extended to approximately 500 ms to allow for reliable noise estimation. The FFT size (N) was 512. The objective evaluation was carried out on the NOIZEUS corpus using the PESQ (perceptual evaluation of speech quality) measure [15].

The treatment types used in the evaluations are listed below (p is the order of the LPC analysis):

1. original clean speech (**Clean**);
2. speech corrupted with white Gaussian noise (**Noisy**);
3. time-domain Kalman filter with LPCs estimated from clean speech, $p = 10$, 40 ms frame duration with 5 ms update, Hamming window (**TDKF ideal**);
4. modulation-domain Kalman filter with LPCs estimated from clean speech, $p = 2$, 10 ms frame duration with 2.5 ms update in modulation domain, (**MDKF ideal**);
5. modulation-domain Kalman filter with LPCs estimated from noisy speech, $p = 2$, 10 ms frame duration with 2.5 ms update in modulation domain, (**MDKF noisy**);
6. modulation-domain Kalman filter with LPCs estimated from MMSE-STSA enhanced speech, $p = 2, q = 4$, 80 ms frame duration with 10 ms update in modulation domain (**MDKF-MMSE**);
7. iterative time-domain Kalman filter [2] with three iterations, $p = 10$, 32 ms frame duration with 4 ms update (**TDKF iterative**);
8. MMSE-STSA method [16] (**MMSE-STSA**); and
9. phase spectrum compensation method [17] (**PSC**).

3.2. Results and discussion

Table 1 shows the average PESQ scores of the modulation-domain Kalman filtering methods as well as other speech enhancement methods. It can be seen that the ideal MDKF, which estimated the LPCs from the modulating signals of clean speech, has achieved the highest PESQ of all the enhancement methods. The ideal MDKF has even outperformed the ideal TDKF, which often serves as an upper bound of enhancement performance. However, when the LPCs were estimated from

Table 1: Average PESQ scores comparing the different speech enhancement methods with the proposed method for speech corrupted by white noise. Bold numbers show the best score.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	1.66	1.84	2.18	2.48
Acoustic and time-domain methods:				
TDKF ideal	2.61	2.84	3.14	3.43
TDKF iterative	2.08	2.46	2.81	3.14
MMSE-STSA	1.98	2.36	2.69	2.98
PSC	1.97	2.35	2.72	3.06
Modulation-domain Kalman filtering:				
MDKF ideal	**3.35**	**3.45**	**3.72**	**3.92**
MDKF noisy	1.70	2.00	2.33	2.65
MDKF-MMSE	2.26	2.56	2.87	3.19

the noisy modulating signals (as in MDKF noisy), we observe a dramatic drop in PESQ, as is normally expected when using poor LPC estimates. When the MMSE-STSA algorithm was used to pre-enhance the speech frame in the acoustic domain prior to LPC analysis in the modulation domain (as in MDKF-MMSE), we can see some improvements in the PESQ scores, where it outperformed all acoustic and time-domain methods except for the ideal TDKF.

When comparing the spectrograms in Figs. 4(c) and 4(f), we notice that the enhanced speech from the MDKF has slightly more detail than the output of the TDKF, which correlates with the higher PESQ scores seen in Table 1. In informal listening tests comparing the ideal MDKF with the TDKF, we found the latter method to produce speech that tended to sound ‘breathy’ and unvoiced, while the former method produced clearer speech. We may attribute this problem of the TDKF to the loss of the fine structure (long-term correlation information), when the Kalman filter weights the linear predictor (which uses only short-term correlations) more favourably over the noisy observation. The MDKF does not suffer from this problem since the Kalman filter is processing in the modulation domain (time trajectories of spectral magnitudes). We can also see that there is less residual noise in the MDKF-MMSE output (Fig. 4(g)) than in the MMSE-STSA and PSC methods in Figs. 4(d) and 4(e).

4. Conclusion

In this paper, we have proposed the use of Kalman filtering in the modulation domain for speech enhancement. In contrast to previous modulation domain-enhancement methods, the modulation-domain Kalman filter (MDKF) is an adaptive MMSE estimator that uses the statistics of temporal changes in magnitude spectrum for both speech and noise. Furthermore, since the modulation phase plays a more important role than acoustic phase, the Kalman filter is highly suited since it is a joint magnitude and phase spectrum estimator, under non-stationarity assumptions. Experimental results from the NOIZEUS corpus showed the MDKF (with clean speech parameters) to outperform all the acoustic and time-domain enhancement methods evaluated.

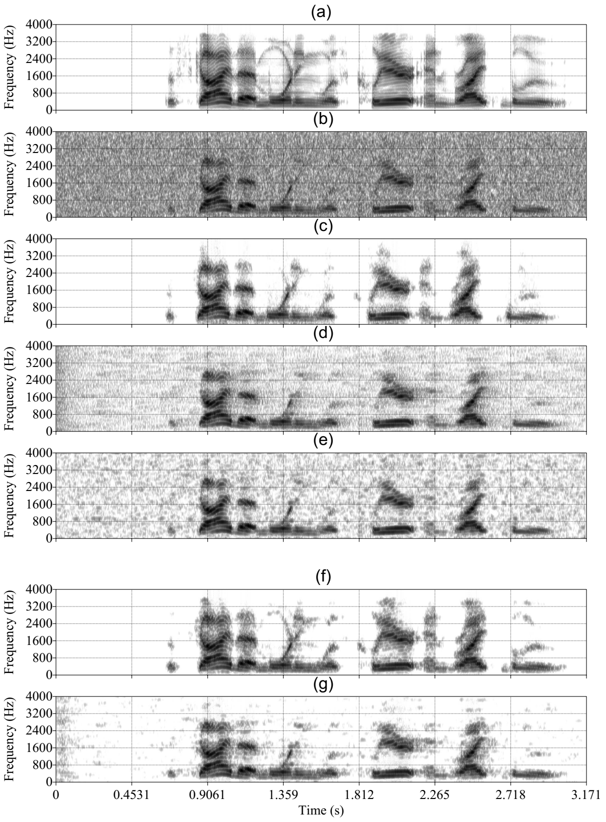


Figure 4: Spectrograms of sp10.wav (*‘The sky that morning was clear and bright blue’*) from the NOIZEUS corpus corrupted with white Gaussian noise at an SNR of 5 dB: (a) Clean speech; (b) noise-corrupted speech (PESQ=1.8); (c) TDKF ideal (PESQ=3.27); (d) MMSE-STSA (PESQ=2.29); (e) PSC (PESQ=2.2); (f) MDKF ideal (PESQ=3.69); (g) MDKF-MMSE (PESQ=2.54).

5. References

[1] 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.

[2] 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.

[3] C. J. Li, “Non-Gaussian, non-stationary, and nonlinear signal processing methods – with applications to speech processing and channel estimation,” Ph.D. dissertation, Aarlborg University, Denmark, Feb. 2006.

[4] P. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. CRC Press LLC, 2007.

[5] L. Atlas and S. A. Shamma, “Joint acoustic and modulation frequency,” *EURASIP Journal of Applied Signal Processing*, vol. 2003, pp. 668–675, Jan. 2003.

[6] N. Mesgarani and S. Shamma, “Speech enhancement based on filtering the spectrotemporal modulations,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2005, pp. 1105–1108.

[7] R. Drullman, J. M. Festen, and R. Plomp, “Effect of temporal envelope smearing on speech reception,” *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[8] ——, “Effect of reducing slow temporal modulations on speech reception,” *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 2670–2680, May 1994.

[9] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, “Syllable intelligibility for temporally filtered lpc cepstral trajectories,” *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2783–2791, 1999.

[10] H. Hermansky, E. Wan, and C. Avendano, “Speech enhancement based on temporal processing,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1995, pp. 405–408.

[11] T. Falk, S. Stadler, W. B. Kleijn, and W. Y. Chan, “Noise suppression based on extending a speech-dominated modulation band,” in *Proc. European Signal Processing Conference*, Aug. 2007, pp. 970–973.

[12] J. G. Lyons and K. K. Paliwal, “Effect of compressing the dynamic range of the power spectrum in modulation filtering based speech enhancement,” in *Proc. INTERSPEECH 2008*, Sep. 2008, pp. 387–390.

[13] K. K. Paliwal, K. K. Wojcicki, and B. Schwerin, “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain,” *Speech Commun.*, vol. 52, no. 5, pp. 450–475, May 2010.

[14] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice-Hall, 2002.

[15] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P.862.” ITU-T, Tech. Rep., 2001.

[16] 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.

[17] K. K. Wojcicki, M. Milacic, A. P. Stark, J. G. Lyons, and K. K. Paliwal, “Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement,” *IEEE Signal Process. Lett.*, vol. 15, pp. 461–464, 2008.