INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Two-Stage Temporal Processing for Single-Channel Speech Enhancement

Suman Samui, Indrajit Chakrabarti, Soumya Kanti Ghosh

Indian Institute of Technology, Kharagpur, India
samuisuman@gmail.com, indrajit@ece.iitkgp.ernet.in, skg@iitkgp.ac.in

Abstract

Most of the conventional speech enhancement methods operating in the spectral domain often suffer from spurious artifact called *musical noise*. Moreover, these methods also incur an extra overhead time for noise power spectral density estimation. In this paper, a speech enhancement framework is proposed by cascading two temporal processing stages. The first stage performs excitation source based temporal processing that involves identifying and boosting the excitation source based speech-specific features present at the gross and fine temporal levels, whereas the second stage provides noise reduction by estimating standard deviation of noise in time-domain by using a robust estimator. The proposed noise reduction stage is quite simply implementable and computationally less complex as it does not require noise estimation in spectral domain as a pre-processing phase. The experimental results have established that the proposed scheme produces on an average 60-65 % improvement in the speech quality (PESQ scores) and intelligibility (STOI scores) at 0 and -5 dB input SNR when compared to existing standard approaches.

Index Terms: Speech enhancement, noise-reduction, noise estimation, temporal processing.

1. Introduction

Despite the presence of enormous research efforts in the last few decades [1], single channel speech enhancement is still considered as one of the most challenging problem, mainly because of two reasons; Firstly, the nature and characteristics of speech signals change frequently in time and from application to application. Secondly, the additive background noise is non-stationary. So, it does not affect the speech spectrum uniformly and it is very difficult to accurately estimate the noise spectrum from spectral analysis [2].

1.1. Motivation

Most of the conventional solutions have been proposed in spectral domain [1][3][4][5], where the noise statistics is estimated from the STFT spectrum of the noisy speech. The magnitude spectrum of desired clean speech is estimated by multiplying a frequency dependent spectral gain function with the noisy signal spectrum. These methods are useful only when the acoustic noise is stationary and SNR is high. Apart from a few recent deep learning based methods of speech separation [6][7], none of the well-known techniques of speech enhancement based on signal processing was found to be promising in improving the speech intelligibility relative to unprocessed corrupted speech [8]. Moreover, these techniques increase the perceived quality at the expense of reduced intelligibility by introducing spurious artefacts known as *musical noise* in the processed speech signal [1]. In the current study, we have presented a speech enhancement system which consists of two temporal processing units in cascaded form. The only input available to the system is the single-channel speech samples corrupted by highly non-stationary acoustic noises. In the first stage, the speech dominated high SNR regions of the given input signal are selectively enhanced by weighting the linear prediction (LP) residual signal samples. A weight function is derived mainly by analyzing the excitation source based speech-specific gross level features (sum of the peaks in the discrete Fourier transform (DFT) spectrum, smoothed Hilbert envelope (HE) of the LP residual and modulation spectrum values) and fine temporal level features i.e. the instants of significant excitation [9][10]. The weighted residual signal samples are used to excite the time-varying all-pole LPC synthesis filter to produce enhanced speech. As this stage does not model the corrupting noise, it is ineffective in minimizing the degrading component and so the output of this first stage contains contaminating noise components. Henceforth, the output signal is subjected to another stage of processing which mainly aims to reduce the noise components from the signal. The second stage also processes the signal on a frame by frame basis in temporal domain and estimates noise standard deviation of each frame using a DATE (d-dimensional trimmed estimator) [11]. These noise standard deviation values are used to define a noise selection threshold in every frame.

1.2. Contribution of the current work

- The main advantage of the proposed approach is that the noise reduction stage (i.e. stage 2) does not require any spectral analysis as a pre-processing stage. So, the proposed method is computationally less complex and faster.
- The proposed approach does not require modelling of noise and it can be applied to enhance the speech for any kind of additive noise.
- Moreover, unlike spectral domain algorithms such as spectral subtraction [12], MMSE-LSA [4] etc, this method is almost free from annoying tonal artefact *musical noise*.
- Although in [11], DATE has been used only for Gaussian noise, but our extensive experiments have demonstrated that the proposed approach yields consistent improvement in terms of PESQ (perceptual evaluation of speech quality) and STOI (short-time objective intelligibility) scores in multiple noise conditions at various SNR levels when compared to other standard speech enhancement methods.

The remainder of the paper is organized as follows. In Section II, the proposed speech enhancement framework has been explained. Section III presents the experimental results. Finally, Section IV concludes the work.

http://dx.doi.org/10.21437/Interspeech.2016-307

2. Proposed framework

The proposed framework of speech enhancement consists of two processing units $TP1$ and $TP2$. The aim of the first stage of temporal processing ($TP1$) is to identify the high SNR regions in the noisy speech by exploiting vocal tract and excitation source informations, and enhance them relative to the low SNR regions while the TP2 provides noise reduction.

2.1. Stage-1 processing (TP1)

TP1 involves detecting and enhancing speech specific features present at the gross and fine temporal levels. TP1 consists of three main steps and mostly follows the processing demonstrated in [13]. The description of TP1 method is summarized as follows:

Step-1 Gross-level processing:
i) As the peaks in the DFT spectrum provide the vocal tract information, the sum of 10 largest amplitude peaks for each frame in STFT domain are computed.
ii) The smoothed Hilbert envelope (HE) of the LP residual is computed which contains the excitation source information.
iii) Next, the modulation spectrum is obtained which gives the long-term (supra-segmental) information of speech production.
iv) These three parameters are enhanced by first order differentiation (FOD) [14]. The summed values of these parameters are normalized w.r.t the maximum value.
v) These normalized sum values are again non-linearly mapped by sigmoid function [13]. Thus one has to compute

$$w_g(n)\frac{1}{1+e^{-\lambda(s_i(n)-T)}} \quad (1)$$

where λ is the slope parameter. T is the average value of the normalized sum $s_i(n)$. These mapped values i.e. $w_g(n)$ are termed as $gross\ weight\ function$ (GWF).
Step-2 Fine temporal level processing:
i) It involves correctly identifying the instants of significant excitation from the given speech utterance. The instants of significant excitation can be obtained from HE of the LP residual. However, HE envelope of the degraded speech spuriously detects wrong peaks as the instants of significant excitation due to the noise components. Therefore, the sinusoidal analysis is done on the noisy speech signal to eliminate most of the noise components. In the sinusoidal modeling, the excitation signal can be represented as the sum of a finite number of sinusoidal parameters. The speech signal is synthesized by taking the eight largest peaks in each frame of STFT analysis.
ii) After that, LP analysis is performed on the synthesized speech signal. Then the HE of the LP residual is computed and mean smoothed using 1 ms rectangular window to smooth out the smaller variations. The peaks in the large error regions, representing the instants of significant excitation are detected using the first order Gaussian differentiator (FOGD) [13]

$$f_d(n)=\frac{1}{\sigma\sqrt{2\pi}}[e^{-}\frac{(n+1)^2}{2\sigma^2}-e^{-}\frac{n^2}{2\sigma^2}], 1\le n\le L_g \quad (2)$$

where the length of the Gaussian window L_g = 80 samples and the standard deviation σ is set to 8.
iii) Convolve the negative of FOGD operator with the mean smoothed HE of the LP residual and determine the negative to positive transitions.
iv) Detected instants are convolved with 3 ms hamming window. The obtained function is termed as $fine\ weight\ function$ (FWF).

Step-3 Synthesis:
i) Combined weight function (CWF) is derived by multiplying two weight functions GWF and FWF.
ii) The LP residual of noisy speech is multiplied with final weight function.
iii) Finally, the processed speech is produced by exciting the all-pole LPC synthesis filter by modified LP residual signal.
Figure 2 depicts an instant of TP1 processing. The derived GWF, FWF, CWF and the output of TP1 are shown in Figure 2(b)-(e).

2.2. Stage-2 processing (TP2)

Algorithm 1 TP2

Inputs: Input noisy signal: $y(n)$.
Output: Output signal: $\hat{y}(n)$.

The input signal is segmented into a set of short-time frames as $y_f(m) = y(m + fM)$, where m=$1, 2...., M$, $f \in 0, 1, 2....., N-1$. N and M indicate the number of frames and and frame-length in samples respectively.
initialize detection threshold:
$\eta(\rho) = 0.5\rho + \frac{1}{\rho}\log(1+\sqrt{1-exp(-\rho^2)})$
Sort out the noisy sequence ($y_f(m)$) in ascending order of their amplitude values i.e. $y_f(1) \le y_f(2) \le y_f(3).... \le y_f(M)$.
Using Bienayme-Chebyshev-Markov inequality (BCMI) compute k_{min}=$M/2 - hM$, where h=$\frac{1}{\sqrt{4M(1-C)}}$ and C denotes the confidence degree.
j:=0 ; p:=k_{min}
for $j \le N-1$ **do**
  for $p \le M$ **do**
    if $||y_j(p-1)|| \le \frac{\eta(\rho)}{\lambda}\frac{\sum_{i=1}^{M}||y_i||}{p} \le ||y_j(p+1)||$ **then**
      $b_f = p$
    else
      $b_f = k_{min}$
    end if
    $\sigma_j = \frac{\eta(\rho)\sum_{i=1}^{b_f}||y_i||}{\lambda b_f}$
    if $y_j(k) \ge y(b_j)$ **then**
      $\tilde{y}_j(p) = y_j(p) - \sigma_j$
    else
      $\tilde{y}_j(p) = 0$
    end if
  end for
end for
Reconstruction: $\tilde{y}(n) = \sum_{j=0}^{N-1}\tilde{y}_j(p-jM)$

The output of the TP1 as shown in Figure 2(e) contains boosted speech components, but still has some noise components as TP1 does not model degrading noise information. Henceforth, the signal obtained from TP1 is further subjected to another temporal processing TP2 which processes the input signal frame by frame basis and estimates noise standard deviation (σ) based on a robust estimator DATE estimator [11]. DATE is one type of trimmed estimator which trims the magnitude or norms of the observed samples by assuming that the signal norms are above some unknown lower bound and that signal probabilities of occurrence are less than one half. Here we have considered one dimensional analysis i.e. d=1. The

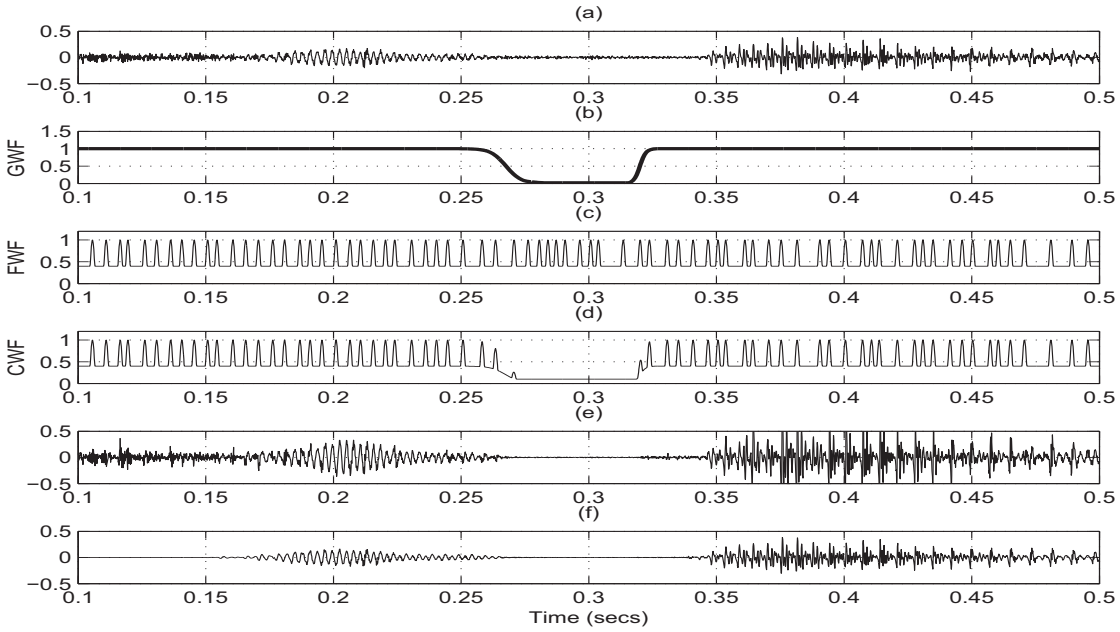

Figure 1: (a) Speech utterance contaminated by babble noise with 0 dB SNR (b) Gross weight function obtained from gross level processing of TP1 (c) Fine weight function obtained from fine temporal level processing of TP1 (d) Output signal of TP1 (e) Combined weight function obtained from fine temporal level processing of TP1 (f) Output of TP2 i.e. the enhanced signal.

values of noise standard deviation are used to find frames of the noisy signal where speech can be considered to be absent. These values are also used to define a noise selection threshold. The amplitude values below this computed threshold values are treated as noise only samples and so these samples are discarded. Noise standard deviation values of each frame are then subtracted from the remaining samples to obtain the amplitudes of the enhanced speech signal. The detailed steps of TP2 is described in Algorithm 1. Two parameters ρ and λ (known as adjustment parameter) are important for initializing the threshold η and computing the noise selection threshold respectively. In TP2, the values of parameters ρ and λ are empirically selected. In [11], it has been shown that in case of Gaussian noise for ρ=4 and λ=0.7979, the performance of the estimator is the best. The confidence degree C is set to 95% as in [11]. In this work we found out the optimal values of ρ for other noise cases as well. It is evident from Figure 2(f) that the output of TP2 is free from degrading noise compared to output signal from TP1. The spectrograms for noisy speech and enhanced speech are also shown in Figure 3. It can be seen that the speech is enhanced not only in the regions where SNR is high but also in the regions of low SNR, associated with unvoiced speech regions. Since the noise reduction step in TP2 operates in time domain, the harmonic structure of the enhanced speech remains restored as almost same like clean speech.

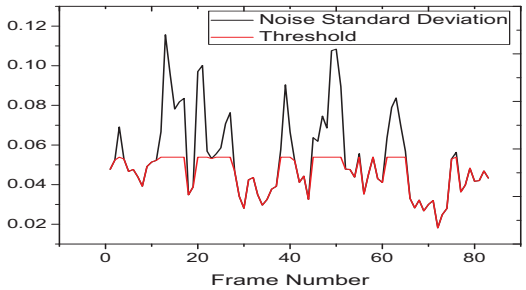

Figure 2: The estimated noise standard deviation and detected threshold using TP2.

3. Experimental results

3.1. Experimental setup and database

To evaluate the performance of the proposed approach, 50 sentences (clean speech samples) are chosen from TIMIT database. The utterances are evenly distributed between male and female speakers. In order to generate noisy stimuli, these clean speech utterances are then contaminated by various non-stationary acoustic noise instances (babble, factory, white Gaussian noise), which have been taken from NOISEX-92 database at various SNR levels (-5 dB to +10 dB) with 5 dB step.

3.2. Speech quality and intelligibility evaluation

Two objective measures PESQ and STOI [15] which are found to be highly correlated with subjective listening tests, are selected for the assessment of perceived speech quality and intelligibility respectively. We computed the PESQ and STOI scores for all the twenty five sentences for different noisy conditions at different SNR levels and compared with standard benchmark methods like MMSE-LSA method [4], multi band spectral subtraction (MBSS) [16], Combined spectral and temporal processing (CSTP) [13], consistent Wiener filtering (CWF) [17], Kalman filtering [18]. The average test result of PESQ and STOI improvement ($\Delta PESQ$ and $\Delta STOI$) for various noise scenarios is reported in Table 1. The best results for every noise conditions are shown in boldface. It is worth noticing that the proposed method outperforms all other methods in terms of

Table 1: PESQ and STOI improvement averaged over test data [represented as Δ PESQ(Δ STOI)]

Type	SNR	Proposed Approach	MMSE-LSA	CSTP	CWF	Kalman Filter
Babble	-5 dB	**0.16(0.06)**	-0.03(-0.06)	0.05(-0.07)	-0.02(-0.05)	0.02(-0.08)
	0 dB	**0.18(0.06)**	-0.06(-0.16)	0.12(-0.06)	0.06(-0.06)	0.06(-0.06)
	5 dB	**0.46(0.29)**	0.06(-0.16)	0.06(-0.06)	0.12(-0.06)	0.19(-0.06)
	10 dB	0.52(0.36)	0.41(0.23)	**0.53(0.26)**	0.39(0.18)	0.36(0.15)
AWGN	-5 dB	**0.18(0.13)**	-0.02(-0.08)	0.04(-0.09)	-0.06(-0.03)	0.02(-0.08)
	0 dB	**0.27(0.16)**	0.17(-0.16)	0.18(-0.06)	0.12(-0.08)	0.11(-0.10)
	5 dB	**0.48(0.32)**	0.26(0.16)	0.36(0.18)	0.12(-0.06)	0.19(-0.06)
	10 dB	**0.65(0.42)**	0.55(0.21)	0.48(0.26)	0.39(0.18)	0.36(0.15)
Factory	-5 dB	**0.23(0.17)**	-0.08(-0.10)	0.04(-0.08)	-0.16(-0.13)	0.12(-0.18)
	0 dB	**0.35(0.19)**	0.11(-0.12)	0.24(-0.08)	0.08(-0.02)	0.16(-0.17)
	5 dB	0.49(0.37)	0.22(0.12)	**0.52(0.36)**	0.19(-0.09)	0.13(-0.16)
	10 dB	**0.68(0.35)**	0.52(0.37)	0.43(0.16)	0.35(0.24)	0.31(0.12)

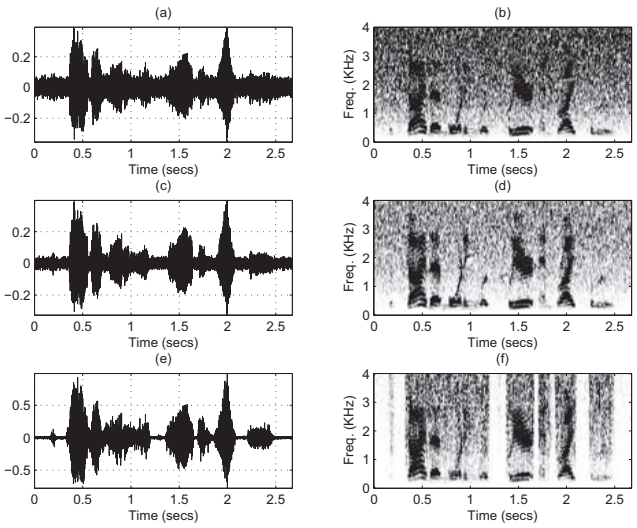

Figure 3: Speech utterance in time domain and corresponding spectrograms: (a, b) spectrogram (c, d) speech waveform processed by only TP1 and its spectrogram ; (e, f) speech waveform processed by proposed approach and its spectrogram.

PESQ and STOI evaluation, even for very low input SNR (-5 dB and 0 dB).

The overall speech quality also has been evaluated by using composite objective measure [19] metric C_{ovl} which has been computed by linearly combining various objective measure metrics such as PESQ, WSS (Weighted spectral slope), LLR (log-likelihood ratio), IS (Itakura-Saito distance), CEP (cepstrum distance), Segmental SNR (SNR_{seg}) and frequency weighted segmental SNR ($fwSNR_{seg}$):

$$\begin{aligned} C_{ovl} =& 1.594 + 0.344 fwSNR_{seg} + 0.805 PESQ \\ & - 0.512 LLR + 0.006 IS + 0.141 CEP - 0.007 WSS \\ & + 0.033 SNR_{seg} \end{aligned} \quad (3)$$

The comparative result of overall signal quality improvement (ΔC_{ovl}) as shown in Figure 4, clearly indicates that the signal quality improvement of our approach is also quite significant compared to other methods.

3.3. Computational complexity

We have compared the relative computational time for the proposed method with other conventional speech enhancement

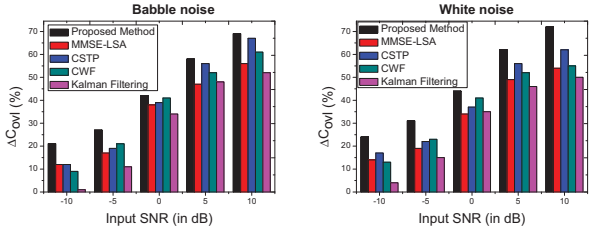

Figure 4: Improvement in overall signal quality.

Table 2: Normalized mean processing time

Speech enhancement methods					
MMSE-LSA	MBSS	CSTP	Kalman Filter	CWF	Proposed method
1.34	1.28	1.07	1.32	1.14	1

methods [4][16][17][18] by computing the processing time required to run the MATLAB programs (R2014a version) in a PC with Intel core i7 processor and 3.10 GHz clock frequency. The computed values of processing times for all these methods are normalized with respect to the processing time of the proposed method. The results presented in Table 2 imply that the proposed method is also computationally faster than the other existing methods.

4. Conclusions

A speech enhancement framework has been proposed, which composed of mainly two stages of signal processing in temporal domain. The first stage relies on the speech production process, rather than noise data modelling. It boosts the signal components in the high SNR regions by identifying the instants of significant excitation, while second stage of processing provides noise reduction by estimating the noise standard deviation in each frame. It has been experimentally demonstrated that the proposed approach yields consistent improvement in perceived speech quality and intelligibility in terms of various objective measures even for signal with low SNR levels for different noise scenarios. As the proposed approach is computationally less complex, its real-time implementation will be a feasible future work for many portable voice communication devices.

5. Acknowledgments

The authors would like to thank Dr. P. Krishnamoorthy for helping us to implement the first stage of temporal processing (TP1).

6. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] P. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, Jan 2011.

[3] R. C. Hendriks, T. Gerkmann, and J. Jensen, "Dft domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, 2013.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[5] S. Samui, I. Chakrabarti, and S. K. Ghosh, "Improved single channel phase-aware speech enhancement technique for low signal-to-noise ratio signal," *IET Signal Processing (DOI:10.1049/iet-spr.2015.0182)*, 2016.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[7] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[8] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7, pp. 588–601, 2007.

[9] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, no. 1, pp. 25–42, 1999.

[10] B. Yegnanarayana, S. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–541.

[11] D. Pastor and F.-X. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences," *Signal Processing, IEEE Transactions on*, vol. 60, no. 4, pp. 1545–1555, 2012.

[12] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[13] P. Krishnamoorthy and S. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication*, vol. 53, no. 2, pp. 154–174, 2011.

[14] ——, "Reverberant speech enhancement by temporal and spectral processing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 2, pp. 253–266, 2009.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2125–2136, 2011.

[16] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE international conference on acoustics speech and signal processing*, vol. 4. Citeseer, 2002, pp. 4164–4167.

[17] J. Le Roux and E. Vincent, "Consistent wiener filtering for audio source separation," *Signal Processing Letters, IEEE*, vol. 20, no. 3, pp. 217–220, 2013.

[18] S. Gannot, "Speech enhancement: Application of the kalman filter in the estimate-maximize (em) framework," in *Speech Enhancement*. Springer, 2005, pp. 161–198.

[19] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.