Implicit Wiener Filtering for Speech Enhancement In Non-Stationary Noise

Rahul Jaiswal and Daniel Romero
Department of Information & Communication Technology
University of Agder, Grimstad, Norway
{*rahul.jaiswal, daniel.romero*}*@uia.no*

***Abstract*—Speech quality is degraded in the presence of background noise, which reduces the quality of experience (QoE) of the end-user and therefore motivates the usage of speech enhancement algorithms. A large number of approaches have been proposed in this context. However most of them have focused on the case where the noise is stationary, an assumption that seldom holds in practice. For instance, in mobile telephony, noise sources with a marked non-stationary spectral signature include vehicles, machines, and other speakers to name a few. On the other hand, the usage of frequency-domain information in existing algorithms for speech enhancement in non-stationary noise environments can be made more effective by leveraging the increased flexibility introduced by implicit Wiener filters, which allow the control of the spectral reconstruction of the speech signal through the adjustment of hyperparameters. To address these limitations, the present paper develops an algorithm that recursively estimates the noise power spectral density and reconstructs the target speech signal in the frequency domain by means of an implicit Wiener filter with judiciously selected hyperparameters. The recursive noise estimation approach relies on the past and the present power spectral values. To evaluate the performance of the speech enhancement algorithm, speech uttered by a male and a female speaker degraded by non-stationary noise produced e.g. by babbling, cars, street noise, trains, restaurants, and airport noise. To this end, the NOIZEUS corpus is used. Objective speech quality measures such as the log-likelihood ratio (LLR), the cepstral distance (CD), and the weighted spectral slope distance (WSS) are evaluated for the enhanced speech signals and compared to the conventional spectral subtraction method. Results demonstrate that the proposed algorithm provides consistent and improved enhancement performance with all tested noise types.**

***Keywords*—Implicit Wiener filtering, Spectral subtraction, Speech enhancement, Noise estimation, Non-stationary noise.**

I. Introduction

Speech is one of the most fundamental means of communication not only among humans but also between humans and machines, as witnessed by the advances in speech recognition and speaker identification [1]. The need for enhancing arises in situations where the speech signal originates in a noisy location or is affected by noise over a communication channel. Voice communication, for instance, over cellular phones typically suffers from the background noise present in cars, trains, restaurants, etc. Speech enhancement (SE) algorithms can therefore be used to improve the quality of speech e.g. in a pre-processing stage of the speech coding system employed by cellular phones [2] or in the speech recognition system for voice dialing [3]. In an air-to-ground communication scenario [4], SE techniques are needed to improve intelligibility, since the pilot's speech is heavily degraded by the so-called cockpit noise. Similarly, impaired listeners wearing hearing aids (or cochlear implant devices) experience extreme difficulty while communicating in noisy conditions. In this setting, SE algorithms can be used to preprocess the noisy speech signal before amplification as it can reduce the listener's fatigue.

Speech enhancement depends on the characteristics of the noise source or interference, the relationship (if any) of the noise to the clean speech signal, and the number of microphones available. The interference could be approximately stationary, as it occurs e.g. with fan noise, or non-stationary, as in the case e.g. of restaurant noise. Suppressing non-stationary noise is more challenging than suppressing stationary noise because its spectral features are constantly changing. Since real-life noise is typically non-stationary, its power and spectral features need to be extracted from the noisy speech signal alone. Noise power estimation is crucial to effective speech enhancement as inaccurate noise estimation can result in musical noise and speech distortion.

Spectral subtraction (SS) is one of the traditional methods used for enhancing speech degraded by additive stationary background noise [5], [6]. The multi-band spectral subtraction algorithm is proposed in [7] to enhance speech corrupted by fan noise. A modified spectral subtraction algorithm is proposed in [8], where the noise spectrum is updated on the basis of a short-term energy measure. These methods perform well in the presence of stationary noise, but suffer from a common limitation in non-stationary environments: they introduce the so-called *musical noise*. Spectral subtraction also does not attenuate noise sufficiently during silence periods.

Wiener filtering (WF) [9] is an alternative method to spectral subtraction for enhancing the speech signal. For additive white Gaussian noise (AWGN) and colored noise, this has been presented in [10]. A technique in the so-called empirical mode decomposition (EMD) domain to enhance the signal corrupted by AWGN noise using the Wiener filter and spectral subtraction is proposed in [11]. These methods, again, work well in the presence of stationary noise. Wavelet denoising [12] is another method based on the wavelet decomposition of noisy signal and thresholding in the wavelet domain to remove background noise and enhancing the speech. However, this method distorts some useful components of the original speech.

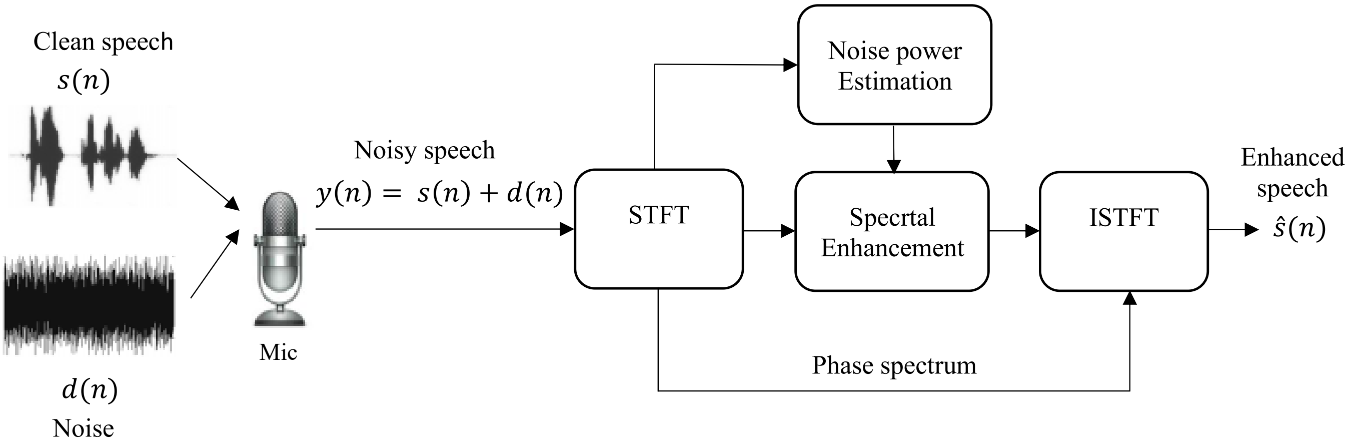

Fig. 1. Block diagram of single-channel speech enhancement system.

In this paper, we propose a simple and reliable single-channel speech enhancement algorithm for non-stationary noise based on the so-called *implicit* Wiener filter [13]. The parameters featuring this algorithm introduce an additional degree of flexibility that allows the engineer to control how an estimate of the noise power spectral density (PSD) is used to suppress the noise component. Relying on a single channel is especially useful in mobile communication applications, where only a single microphone is available due to the cost and size reasons. The algorithm is fed with an estimate of the noise PSD obtained by means of a running average of the noise spectral energy.

The block diagram of the proposed single-channel speech enhancement system is shown in Fig. 1. Its performance is extensively compared with the aforementioned spectral subtraction method on a real data set that comprises clean and corrupted speech signals.

The remainder of this paper is laid out as follows: Section II presents the review of the spectral subtraction method and Section III describes the implicit Wiener filter in the frequency domain. Section IV presents the noise power spectral density (PSD) estimation approach. Section V describes the experimental dataset and Section VI outlines the evaluation methodology of speech enhancement algorithms. Section VII presents and discusses the simulation results. Finally, concluding remarks and future directions are presented in Section VIII.

II. The Spectral Subtraction Method

Spectral subtraction (SS) [5], [14] is one of the most popular methods for single-channel speech enhancement. The basic principle of SS is to obtain an estimate of the clean speech signal spectrum by simply subtracting the noise spectrum from the noisy speech spectrum. It relies on the following assumptions. Firstly, the speech signals are assumed stationary and the noise spectrum must not change abruptly in between frames. Secondly, the speech signals are degraded by statistically uncorrelated and independent additive noise with zero mean. Finally, it is assumed that the phase distortion is not perceived by the human ear. For this reason, the phase of noisy speech is kept unchanged in the enhancement stage.

Suppose that the noisy speech signal $y(n)$ can be expressed in terms of the clean speech signal $s(n)$ and additive noise $d(n)$ as,

$$y(n) = s(n) + d(n) \tag{1}$$

Since this algorithm operates on a frame-by-frame basis, it is convenient to express (1) as,

$$y(n,k) = s(n,k) + d(n,k) \tag{2}$$

where $n = 0, 1, 2, \ldots, (N-1)$ is the discrete-time index, $k = 1, 2, \ldots$ is the frame number, and N is the length of the frame. Taking the discrete-time Fourier transform of both sides with respect to n yields

$$Y(\omega,k) = S(\omega,k) + D(\omega,k) \tag{3}$$

where ω is the discrete angular frequency. To obtain the short-time power spectrum of $y(n)$, we multiply both sides of (3) by their complex conjugates, which yields

$$\begin{aligned}|Y(\omega,k)|^2 &= |S(\omega,k)|^2 + |D(\omega,k)|^2 + S(\omega,k)D^*(\omega,k) + \\ &\quad D(\omega,k)S^*(\omega,k) \\ &= |S(\omega,k)|^2 + |D(\omega,k)|^2 + 2\mathrm{Re}\{S(\omega,k)D^*(\omega,k)\}\end{aligned} \tag{4}$$

The terms $|D(\omega,k)|^2$, $S(\omega,k)D^*(\omega,k)$ and $D(\omega,k)S^*(\omega,k)$ can not be obtained directly and thus are approximated as $\mathbb{E}[|D(\omega,k)|^2]$, $\mathbb{E}[S(\omega,k)D^*(\omega,k)]$ and $\mathbb{E}[D(\omega,k)S^*(\omega,k)]$ respectively, where $\mathbb{E}[\cdot]$ denotes the expectation operator.

Typically, $\mathbb{E}[|D(\omega,k)|^2]$ is estimated during non-speech activity. Its estimate is denoted by $\hat{P}_{dd}(\omega,k)$. Due to the assumption of zero-mean noise which is uncorrelated with the clean speech signal, the terms $\mathbb{E}[S(\omega,k)D^*(\omega,k)]$ and $\mathbb{E}[D(\omega,k)S^*(\omega,k)]$ reduce to zero. Thus, an estimate of the clean speech power spectrum can be obtained as,

$$\hat{P}_{ss}(\omega,k) = P_{yy}(\omega,k) - \hat{P}_{dd}(\omega,k) \tag{5}$$

where $\hat{P}_{ss}(\omega,k)$ is the enhanced speech power spectrum and $P_{yy}(\omega,k) = |Y(\omega,k)|^2$ is the noisy speech power spectrum. The enhanced speech signal is then obtained by computing the inverse Fourier transform of the square root of $\hat{P}_{ss}(\omega,k)$, using the phase of $Y(\omega,k)$.

The major drawback of the SS method is that the spectrum of the enhanced speech signal obtained using (5) may contain negative values. As a result, a "new" noise appears in the processed speech signal which has been described as ringing or warbling with tonal quality. This is referred to as "musical noise" [14] and affects the human listening.

III. The Implicit Wiener Filter

As mentioned earlier, the enhanced speech signal computed by the spectral subtraction method is highly affected by musical noise. Thus, we turn our attention to the Wiener filter, which is conceptually similar to spectral subtraction but replaces the direct subtraction with an estimate of the clean speech signal spectrum obtained by minimizing the mean square error [14], [15].

In the conventional Wiener filter, the output signal $\hat{s}(n)$ is obtained as the convolution of the two-sided filter impulse response $h(n)$ and the input signal $y(n)$:

$$\hat{s}(n) = \sum_{k=-\infty}^{\infty} h_k y(n-k) = h(n) \circledast y(n) \tag{6}$$

where $\circledast$ denotes convolution. Therefore, in the frequency domain, one can write,

$$\hat{S}(\omega) = H(\omega)Y(\omega) \tag{7}$$

where $\hat{S}(\omega)$, $H(\omega)$ and $Y(\omega)$ are the discrete-time Fourier transforms of $\hat{s}(n)$, $h(n)$ and $y(n)$ respectively. The estimation error is given by

$$E(\omega) = S(\omega) - \hat{S}(\omega) = S(\omega) - H(\omega)Y(\omega) \tag{8}$$

The Wiener filter finds the $H(\omega)$ that minimizes the mean square error. Upon multiplying both sides of (8) by their complex conjugates, the mean square error is given by

$$\begin{aligned}\mathbb{E}[|E(\omega)|^2] &= \mathbb{E}[[S(\omega) - H(\omega)Y(\omega)][S(\omega) - H(\omega)Y(\omega)]^*] \\ &= \mathbb{E}[|S(\omega)|^2] - H^*(\omega)\mathbb{E}[Y^*(\omega)S(\omega)] - \\ &\quad H(\omega)\mathbb{E}[S^*(\omega)Y(\omega)] + |H(\omega)|^2\mathbb{E}[|Y(\omega)|^2]\end{aligned} \tag{9}$$

By letting $P_{yy}(\omega) = \mathbb{E}[|Y(\omega)|^2]$ denotes the power spectrum of $y(n)$ and $P_{sy}(\omega) = \mathbb{E}[S(\omega)Y^*(\omega)]$ the cross-power spectrum of $y(n)$ and $s(n)$, one can express (9) as

$$\begin{aligned}J = \mathbb{E}[|E(\omega)|^2] = \mathbb{E}[|S(\omega)|^2] - 2\mathrm{Re}\{H^*(\omega)P_{sy}(\omega)\} \\ +|H(\omega)|^2 P_{yy}(\omega)\end{aligned} \tag{10}$$

To find the optimal $H(\omega)$, one can take the complex derivative of the mean square error or cost function J with respect to $H(\omega)$[1] and set it equal to zero:

$$\frac{\partial J}{\partial H} = 0 - 2P_{sy}(\omega) + 2H(\omega)P_{yy}(\omega) = 0 \tag{11}$$

$$\left[H(\omega) = \frac{P_{sy}(\omega)}{P_{yy}(\omega)}\right] \tag{12}$$

This expression provides the transfer function of the Wiener filter. To evaluate (12), one needs to compute $P_{sy}(\omega)$ and $P_{yy}(\omega)$. On the one hand,

$$\begin{aligned}P_{sy}(\omega) &= \mathbb{E}[Y^*(\omega)S(\omega)] \\ &= \mathbb{E}[\{S(\omega) + D(\omega)\}^* S(\omega)] \\ &= \mathbb{E}[|S(\omega)|^2] + \mathbb{E}[D^*(\omega)S(\omega)] \\ &= \mathbb{E}[|S(\omega)|^2] \\ &= P_{ss}(\omega)\end{aligned} \tag{13}$$

Here, the fourth equality follows from the fact that the noise is zero-mean and uncorrelated with the clean speech signal, which implies that the cross-term $\mathbb{E}[D^*(\omega)S(\omega)]$ reduces to zero. The fifth equality is a definition.

On the other hand,

$$\begin{aligned}P_{yy}(\omega) &= \mathbb{E}[Y(\omega)Y^*(\omega)] \\ &= \mathbb{E}[\{S(\omega) + D(\omega)\}\{S(\omega) + D(\omega)\}^*] \\ &= \mathbb{E}[|S(\omega)|^2] + \mathbb{E}[S(\omega)D^*(\omega)] + \mathbb{E}[D(\omega)S^*(\omega)] \\ &+ \mathbb{E}[|D(\omega)|^2] \\ &= P_{ss}(\omega) + P_{dd}(\omega)\end{aligned} \tag{14}$$

where $P_{ss}(\omega)$ and $P_{dd}(\omega)$ are respectively defined as $P_{ss}(\omega) = \mathbb{E}[|S(\omega)|^2]$ and $P_{dd}(\omega) = \mathbb{E}[|D(\omega)|^2]$.

Finally, substituting Equations (13) and (14) in Equation (12), the transfer function of Wiener filter reads as

$$\left[H_{WF}(\omega) = \frac{P_{ss}(\omega)}{P_{ss}(\omega) + P_{dd}(\omega)}\right] \tag{15}$$

Here, $H_{WF}(\omega)$ is real, non-negative, and even because $P_{dd}(\omega) \geq 0$ and $P_{ss}(\omega) \geq 0$. Also, $0 \leq H_{WF}(\omega) \leq 1$. This implies that the impulse response $h(n)$ must be even as well, resulting in a non-causal impulse response $h(n)$. Therefore, the Wiener filter is not realizable and can not be applied directly to estimate $s(n)$ in the time domain. For this reason, the proposed algorithm will operate in the frequency domain.

So far, we have assumed that $P_{ss}(\omega)$ and $P_{dd}(\omega)$ are known. However, in practice they must be estimated and, therefore, the reliability of their estimates is highly dependent on the application setup. For this reason, it is desirable to introduce additional flexibility to control how much the enhancement algorithm relies on these estimates. The so-called modified or parametric Wiener filter [13] achieves this aim by introducing two adjustable parameters β and γ as follows:

$$H(\omega) = \left[\frac{P_{ss}(\omega)}{P_{ss}(\omega) + \gamma P_{dd}(\omega)}\right]^{\beta} \tag{16}$$

Here, β is referred to as the noise suppression factor. If β and γ are both equal to one, (16) reduces to (15).

Furthermore, to accommodate the non-stationarity of the speech signal, it is convenient to introduce the following approximation [13]:

$$P_{ss}(\omega) \approx |\hat{S}(\omega)|^2 \tag{17}$$

[1] $H(\omega)$ is complex valued here because the cross-power spectrum $P_{sy}(\omega)$ is generally complex.

that is, we have approximated the true power spectral density of $s(n)$ by its spectral energy. From (7), the output of the Wiener filter in the frequency domain is given by

$$\hat{S}(\omega) = H(\omega)Y(\omega) \tag{18}$$

Substituting (17) into (16) and the resulting expression in (18) yields the following implicit estimator [13]

$$\hat{S}(\omega) = \left[\frac{|\hat{S}(\omega)|^2}{|\hat{S}(\omega)|^2 + \gamma P_{dd}(\omega)} \right]^{\beta} Y(\omega) \tag{19}$$

Clearly, the phase of $\hat{S}(\omega)$ must equal that of $Y(\omega)$. Therefore, one just needs to equate the magnitude of both sides of (19).

For illustration purposes, we now describe how (19) can be solved when $\beta = 1/2$. In this case, squaring both sides of (19) results in

$$|\hat{S}(\omega)|^2 = \left[\frac{|\hat{S}(\omega)|^2}{|\hat{S}(\omega)|^2 + \gamma P_{dd}(\omega)} \right] |Y(\omega)|^2 \tag{20}$$

Solving (20), two solutions arise:

$$|\hat{S}(\omega)| = 0 \tag{21a}$$

$$|\hat{S}(\omega)| = \left[|Y(\omega)|^2 - \gamma P_{dd}(\omega) \right]^{1/2} \tag{21b}$$

A solution for $|\hat{S}(\omega)|$ consistent with Equation (20) is Equation (21b) for positive values under the radical. Finally, the enhanced speech signal is estimated as,

$$\hat{s}(n) = \text{IFFT}\left[\hat{S}(\omega) \right] \tag{22}$$

The proposed algorithm operates on a frame-by-frame basis and uses the overlap-add method [16] to recombine the spectra of the individual frames using the phase of $Y(\omega, k)$.

IV. Adaptive Noise PSD estimation

Estimating noise power across frequency is of vital importance as different frequencies of the speech signal are affected by noise more to a different extent [14]. In other words, each spectral component will typically have a different *effective SNR*. But the distribution of the noise energy in the frequency domain also depends on the kind of source and therefore needs to be estimated. For example, most of the energy of noise produced by cars is concentrated in the low frequency range, whereas train and restaurant noise occupy a wider frequency range [14].

On the other hand, as mentioned in the introduction, many sources produce noise with time-varying spectral characteristics. This is the case for example when multiple people speak in the background or when vehicles are passing by. For this reason, suppressing non-stationary noise is more challenging than suppressing stationary noise.

Furthermore, in view of the previous section, it is clear that the quality of enhanced speech depends on the accuracy of the noise PSD estimate. This is because low noise estimates give rise to noisy enhanced signals, whereas high estimates lead to intelligibility loss [17]. When estimating the PSD of the noise, there is a trade-off involving how fast the estimates are adapted to changes. On the one hand, if the PSD of the observations are averaged over longer time windows, the estimates will have a lower variance but they will not track rapid changes in the noise spectrum.

To balance these effects, the proposed algorithm estimates the noise PSD using the following first-order recursion [14]:

$$\hat{P}_{dd}(\omega, k) = \alpha \hat{P}_{dd}(\omega, k-1) + (1-\alpha) P_{yy}(\omega, k) \tag{23}$$

where α ($0 \le \alpha \le 1$) is the smoothing parameter, k is the frame index, ω is the frequency bin index, $P_{yy}(\omega, k)$ is the short-time power spectrum of the noisy speech signal defined in Sec. II, and $\hat{P}_{dd}(\omega, k)$ is the noise power spectrum estimate in the ω^{th} frequency bin of the k-th frame.

V. Experimental Corpus

To investigate the performance of speech enhancement algorithms in noisy environments, a noisy dataset is needed. NOIZEUS [18] is a publicly available noisy speech corpus, used to facilitate the comparison of speech enhancement algorithms. These are 30 phonetically-balanced IEEE English sentences, spoken by 3 male and 3 female speakers. The sentences are each corrupted with one of six commonly occurring real-world noises: babble, car, street, train, restaurant, and airport at SNRs: 0dB, and 5dB. The noises are taken from the AURORA database. The sentences were originally sampled at 25 kHz and then down-sampled to 8 kHz. The average duration of each utterance is 3 seconds. All sample files are saved in WAV format (16 bit PCM, mono).

VI. Evaluating Speech Enhancement Algorithm

For the evaluation of speech enhancement algorithms, the noisy speech samples from the publicly available NOIZEUS corpus are taken. A total of two phonetically-balanced utterances, one pronounced by a male speaker and one pronounced by a female speaker, are used. The male utterance is "*A good book informs of what we ought to know*" and the female utterance is "*Let us all join as we sing the last chorus*". The noises used are: babble, car, street, train, restaurant and airport at SNRs i.e., 0dB, and 5dB. The speech database is sampled at 8 kHz and quantized linearly using 16 bits resolution. The noise samples used are of zero-mean and the energy of the noisy speech samples are normalized to unity. The frame size is chosen to be 200 samples (25 ms duration), with 50 % overlapping. The sinusoidal Hamming window with size 200 samples is applied to each frame individually. The windowed speech frame is then analysed using Fast Fourier Transform (FFT) with length 256 samples.

The noise is estimated from the noisy speech using the first order recursive Equation (23). As each noise signal has different time-frequency distribution and spectral characteristics, they have a different impact on the speech signal. To get the optimal value of smoothing parameter α to estimate noise of each noise type, the segmental SNR (SNRseg.) is calculated for speech uttered by each speaker (male and female). Segmental SNR varies frame-to-frame in proportion to the signal energy. In the implicit Wiener filtering technique, we considered initial 5 frames of noisy speech as noise/silence to estimate the noise PSD using first order recursive Equation (23), and then used a simple voice activity detector (VAD) to update the noise PSD.

Subjective listening test such as Absolute Category Rating (ACR) [19] is the most reliable method for evaluating speech quality, where a number of people listens the speech samples and rate the quality. However, these tests are costly, time-consuming and impractical for real-time scenarios. As an alternative objective speech quality measures are utilized which are lower cost, fast and practical.

To compare the performance of speech enhancement algorithms, three different objective speech quality measures namely; Log-likelihood ratio (LLR), Cepstral distance (CD), and Weighted spectral slope distance (WSS) are computed. The LLR [14] is the spectral distance measure which mainly models the mismatch between the formants of the clean and the enhanced speech signal. The mean LLR value is obtained by averaging the individual frame LLR values across the sentence. Its value is limited in the range of $[0, 2]$ and it is computed as,

$$d_{LLR}(a_s, \bar{a}_{\hat{s}}) = \log_{10}\left(\frac{\bar{a}_{\hat{s}}^T R_s \bar{a}_{\hat{s}}}{a_s^T R_s a_s}\right) \tag{24}$$

where a_s^T, and $\bar{a}_{\hat{s}}^T$, are the linear prediction coefficients (LPC) of the clean and enhanced speech signal respectively. R_s is the auto-correlation matrix of the clean speech signal. LLR measure is always positive.

The CD [14] provides an estimate of the log spectral distance between two spectra. Its value is limited in the range of $[0, 10]$ and it is computed as,

$$d_{CD}(c_s, \bar{c}_{\hat{s}}) = \frac{10}{\log_e 10}\sqrt{2\sum_{k=1}^{p}\left[c_s(k) - c_{\hat{s}}(k)\right]^2} \tag{25}$$

where $c_s(k)$ and $\bar{c}_{\hat{s}}(k)$ are the cepstrum coefficients (obtained from LPC) of the clean and the enhanced speech signal respectively, and p is the maximum order of the LPC coefficients.

The WSS distance [14] is based on the weighted difference between the spectral slopes in each band. It penalizes heavily difference in spectral peak (formants) locations. Its value is limited in the range of $[0, 150]$ and it is computed as,

$$d_{WSS}(C_s, \bar{C_{\hat{s}}}) = \sum_{k=1}^{36} W(k)\left[S_s(k) - \bar{S}_{\hat{s}}(k)\right]^2 \tag{26}$$

where $S_s(k)$ and $\bar{S}_{\hat{s}}(k)$ are the spectral slopes of the clean and enhanced speech signal of the k^{th} band respectively. $W(k)$ is the weight of the band k.

VII. Simulation Results and Discussions

Table I, Table II, Table III, Table IV, Table V, and Table VI show the segmental SNR (SNRSeg.) of spectral subtraction with recursive noise estimation method for babble, car, street, train, restaurant and airport noise at SNRs 0dB and 5dB with different values of smoothing parameter α, for the speech uttered by both speakers (male and female) respectively.

TABLE I
Segmental SNR of spectral subtraction with recursive noise estimation method for babble noise at different SNR levels.

$\alpha \rightarrow$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SNR ↓										
Male										
0dB	6.228	6.086	6.003	5.946	5.905	5.874	5.850	5.847	**5.872**	5.653
5dB	6.424	6.322	6.274	6.253	6.244	6.243	6.252	6.283	**6.329**	6.227
Female										
0dB	6.494	6.388	6.323	6.275	6.237	6.213	6.200	6.210	**6.246**	6.005
5dB	6.691	6.612	6.566	6.534	6.515	6.508	6.517	6.556	**6.621**	6.666

TABLE II
Segmental SNR of spectral subtraction with recursive noise estimation method for car noise at different SNR levels.

$\alpha \rightarrow$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SNR ↓										
Male										
0dB	6.188	6.045	5.964	5.912	5.880	5.861	5.867	5.884	**5.928**	5.558
5dB	6.412	6.309	6.258	6.232	6.226	6.237	6.266	6.324	**6.419**	6.210
Female										
0dB	6.473	6.333	6.248	6.192	6.157	6.140	6.138	6.157	**6.200**	5.868
5dB	6.695	6.614	6.568	6.544	6.535	6.541	6.566	6.615	**6.718**	6.515

TABLE III
Segmental SNR of spectral subtraction with recursive noise estimation method for street noise at different SNR levels.

$\alpha \rightarrow$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SNR ↓										
Male										
0dB	6.194	6.074	6.003	5.956	5.926	5.918	5.931	5.959	**6.013**	5.618
5dB	6.465	6.371	6.328	6.308	6.305	6.318	6.358	6.422	**6.495**	6.370
Female										
0dB	6.574	6.531	6.499	6.474	6.459	6.455	6.464	6.499	**6.567**	6.422
5dB	6.818	6.767	6.743	6.733	6.739	6.754	6.785	6.844	**6.933**	6.931

TABLE IV
Segmental SNR of spectral subtraction with recursive noise estimation method for train noise at different SNR levels.

$\alpha \rightarrow$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SNR ↓										
Male										
0dB	6.227	6.103	6.032	5.980	5.938	5.908	5.888	5.884	**5.904**	5.595
5dB	6.399	6.306	6.261	6.237	6.226	6.230	6.247	6.284	**6.324**	6.274
Female										
0dB	6.479	6.357	6.280	6.227	6.192	6.175	6.173	6.195	**6.244**	5.918
5dB	6.741	6.655	6.606	6.578	6.567	6.574	6.599	6.647	**6.727**	6.676

From the extensive study of each Table, it can be observed that for every case of input SNR, as the value of α increases then the value of SNRseg. becomes better for each noise type. However, the value of SNRseg. decreases at the extreme end i.e., $\alpha = 1$. As described in Equation (23) that the noise estimation in the current frame is heavily dependent on the noise present in the previous frame as well as lightly dependent on the noisy speech in the current frame. Therefore, from the different values of α, shown in Table I to Table VI, $\alpha = 0.9$ is the best suitable value for our speech enhancement algorithm.

TABLE V

Segmental SNR of spectral subtraction with recursive noise estimation for restaurant noise at different SNR levels.

$\alpha \rightarrow$ SNR $\downarrow$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Male										
0dB	6.380	6.254	6.179	6.129	6.100	6.096	6.095	6.095	**6.117**	5.913
5dB	6.524	6.411	6.349	6.316	6.302	6.301	6.303	6.323	**6.334**	6.235
Female										
0dB	6.694	6.613	6.560	6.519	6.484	6.459	6.444	6.433	**6.466**	6.257
5dB	6.807	6.740	6.704	6.675	6.657	6.649	6.654	6.686	**6.730**	6.662

TABLE VI

Segmental SNR of spectral subtraction with recursive noise estimation method for airport noise at different SNR levels.

$\alpha \rightarrow$ SNR $\downarrow$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Male										
0dB	6.411	6.307	6.248	6.207	6.183	6.159	6.147	6.154	**6.147**	5.770
5dB	6.526	6.458	6.435	6.446	6.467	6.508	6.565	6.651	**6.746**	6.434
Female										
0dB	6.656	6.539	6.465	6.418	6.390	6.377	6.379	6.39 5	**6.454**	6.211
5dB	6.675	6.611	6.577	6.559	6.556	6.576	6.613	6.682	**6.791**	6.633

The Fig. 2 shows the comparison of noisy PSD and the estimated noise PSD using SS and WF with recursive noise estimation algorithm for (a) the speech uttered by a male speaker and degraded by the airport noise at 5dB, and (b) the speech uttered by a female speaker and degraded by the train noise at 5dB. It can be noticed that the noise estimation using the Wiener filter with recursive noise estimation algorithm is performing better than the spectral subtraction. The envelop of the estimated noise PSD matches better with the envelop of the noisy speech PSD using Wiener filter based algorithm. This shows that the Wiener filter with recursive noise estimation algorithm exhibits superior in noise estimation.

Table VII and Table VIII presents the LLR, CD and WSS of the enhanced speech signals using SS and WF algorithms for each type of noise at input SNRs 0dB, and 5dB, for the speech uttered by a male and a female speaker respectively. It can be observed from Table VII, where the utterance is pronounced by a male speaker that the LLR, CD and WSS of wiener filter based speech enhancement algorithm is better than the spectral subtraction for each noise types at each input SNRs, expect for street and airport noise at 0dB, where SS is performing better. This shows that speech enhanced by SS reflects severe perceptual dissimilarity, resulting in very poor noise reduction. The LLR and CD of WF, in case of, airport noise at 5dB are lowest. Similarly, the WSS of WF are lowest, in case of, all noise types as compared to SS. This reflects that WF exhibits the best performance in reducing the noise and airport noise at 5dB, in particular, is reduced significantly higher among all noise types.

TABLE VII

Log-likelihood ratio (LLR), Cepstral distance (CD) and Weighted spectral slope distance (WSS) of enhanced speech signals through Wiener filter (WF) and Spectral subtraction (SS) at 0dB, and 5dB SNRs. English sentence *"A good book informs of what we ought to know"*, pronounced by a male speaker, is used as original signal.

Noise Type	Input SNR	LLR		CD		WSS	
	(dB)	WF	SS	WF	SS	WF	SS
Babble	0	1.181	1.194	6.303	6.281	89.148	102.319
	5	0.980	1.034	5.570	5.732	78.229	93.772
Car	0	0.970	1.046	5.416	5.785	77.215	88.907
	5	0.980	1.034	5.570	5.732	66.670	82.618
Street	0	0.987	0.904	5.569	5.350	73.434	87.803
	5	1.034	1.147	5.861	6.316	75.438	96.333
Train	0	1.406	1.457	7.208	7.609	76.100	90.796
	5	1.230	1.440	6.806	7.844	66.289	89.230
Restaurant	0	1.105	1.131	6.106	6.330	81.515	96.809
	5	0.931	1.096	5.440	6.085	78.248	95.365
Airport	0	0.998	0.961	5.836	5.629	82.857	98.646
	5	0.773	0.957	4.940	5.631	82.002	91.487

TABLE VIII

Log-likelihood ratio (LLR), Cepstral distance (CD) and Weighted spectral slope distance (WSS) of enhanced speech signals through Wiener filter (WF) and Spectral subtraction (SS) at 0dB, and 5dB SNRs. English sentence *"Let us all join as we sing the last chorus"*, pronounced by a female speaker, is used as original signal.

Noise Type	Input SNR	LLR		CD		WSS	
	(dB)	WF	SS	WF	SS	WF	SS
Babble	0	0.956	0.928	5.888	5.680	100.781	118.641
	5	0.933	0.961	5.584	5.772	75.049	101.960
Car	0	0.955	0.991	5.744	5.773	90.319	104.602
	5	0.806	0.841	5.122	5.301	65.674	99.028
Street	0	0.975	1.028	5.680	6.110	78.476	114.885
	5	0.833	0.981	5.107	5.841	69.876	104.651
Train	0	1.041	1.060	6.113	6.254	82.913	104.165
	5	0.753	1.160	4.987	6.469	68.508	91.052
Restaurant	0	0.928	0.995	5.806	5.996	101.839	126.184
	5	0.801	0.957	5.024	5.566	68.720	104.282
Airport	0	0.853	0.935	5.506	5.641	94.243	118.828
	5	0.822	0.805	5.195	5.199	70.526	104.461

It can also be observed from Table VIII, where the utterance is pronounced by a female speaker that the LLR, CD and WSS of wiener filter based speech enhancement algorithm is better than the SS for each noise types at each input SNRs, expect for babble noise at 0dB and airport noise at 5dB, where SS is performing good. This shows that speech enhanced by SS reflects very high perceptual dissimilarity. The LLR and CD of WF, in case of, train noise at 5dB are lowest. Similarly, the WSS of WF are lowest, in case of, all noise types as compared to SS. This reflects that WF exhibits the best performance in reducing the noise and train noise at 5dB, in particular, is reduced significantly higher among all noise types.

While comparing the highly reducible noise (airport noise at 5dB uttered by a male speaker and train noise at 5dB uttered by a female speaker) by the WF algorithm, it can be observed that the LLR and CD of the female uttered sentence is better (lower) than the male uttered sentence. However, the WSS is following the opposite behaviour. This shows that the female uttered sentence, degraded in the presence of airport and train noise, is enhanced more accurately than the male uttered sentence degraded in the same background noise. Further, with the informal listening test, we found that the enhanced speech with the perceptual Wiener filter is more pleasant.

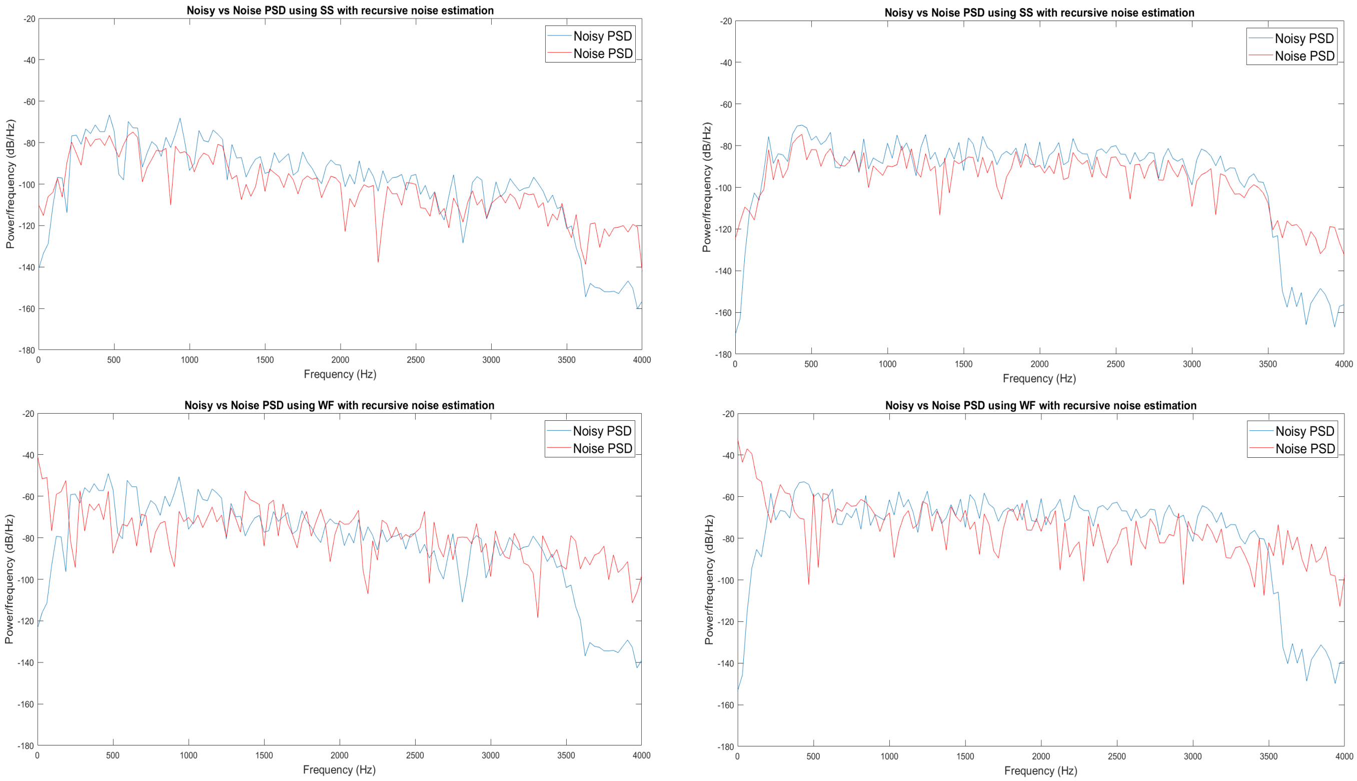

Fig. 2. Noisy vs Noise PSD using SS and WF with recursive noise estimation algorithm for (a) speech uttered by a male speaker and degraded by the airport noise at 5dB (left), and (b) speech uttered by a female speaker and degraded by the train noise at 5dB (right).

The Fig. 3 and Fig. 4 show the time domain representation of the clean speech, noisy speech, enhanced speech using SS and enhanced speech using WF and its corresponding spectrograms for speech uttered by a male speaker and degraded by the airport noise at 5dB, and for speech uttered by a female speaker and degraded by the train noise at 5dB respectively. It can be visualised from the both signal plots that the speech enhanced by the Wiener filter with recursive noise estimation algorithm has better estimation, showing superior performance than the spectral subtraction. Moreover, the spectrograms show that the enhanced speech using Wiener filter has better signal improvement than the spectral subtraction.

VIII. Conclusions and Future Work

In this paper, we have addressed the problem of speech signal estimation which is degraded by the non-stationary noise. The noise power spectral density is estimated using the first order recursive equation and is updated continuously in each frame using a smoothing parameter. The optimal value of smoothing parameter is calculated based on the estimated segmental SNR in each frequency bin of the noisy speech spectrum. The implicit Wiener filter with recursive noise estimation algorithm is proposed to estimate clean speech from the noisy speech and compared to the conventional spectral subtraction method. Results shows that the envelop of the estimated noise using the implicit Wiener filter is quite close to the envelop of noisy speech spectrum as compared to the spectral subtraction. The proposed algorithm yields the enhanced speech signal perceptually similar to the clean speech signal and its spectrogram is also close to the spectrogram of clean speech signal. The musical noise is less structured than the spectral subtraction, while the distortion of the speech remains acceptable. Future work will investigate the integration of different noise estimation techniques to further improve the performance of our algorithm.

References

[1] M. Kolbœk, Z.-H. Tan, and J. Jensen, “Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 305–311.

[2] T. Bäckström, *Speech coding: with code-excited linear prediction*. Springer, 2017.

[3] A. H. Moore, P. P. Parada, and P. A. Naylor, “Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures,” *Computer Speech & Language*, vol. 46, pp. 574–584, 2017.

[4] Z. Zhang, Y. Shi, G. Jia, and J. Yang, “The Comparison of Denoising Methods Based on Air-ground Speech of Civil Aviation,” in *Chinese Conference on Biometric Recognition*. Springer, 2015, pp. 480–487.

[5] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[6] T. Bao, Y. Li, K. Xu, Y. Wang, and W. Hu, “An improved endpoint detection algorithm based on improved spectral subtraction with multi-taper spectrum and energy-zero ratio,” in *International Conference on Intelligent Computing*. Springer, 2018, pp. 266–275.

[7] J. C. Saldanha and O. Shruthi, “Reduction of noise for speech signal enhancement using Spectral Subtraction method,” in *IEEE International Conference on Information Science (ICIS)*, 2016, pp. 44–47.

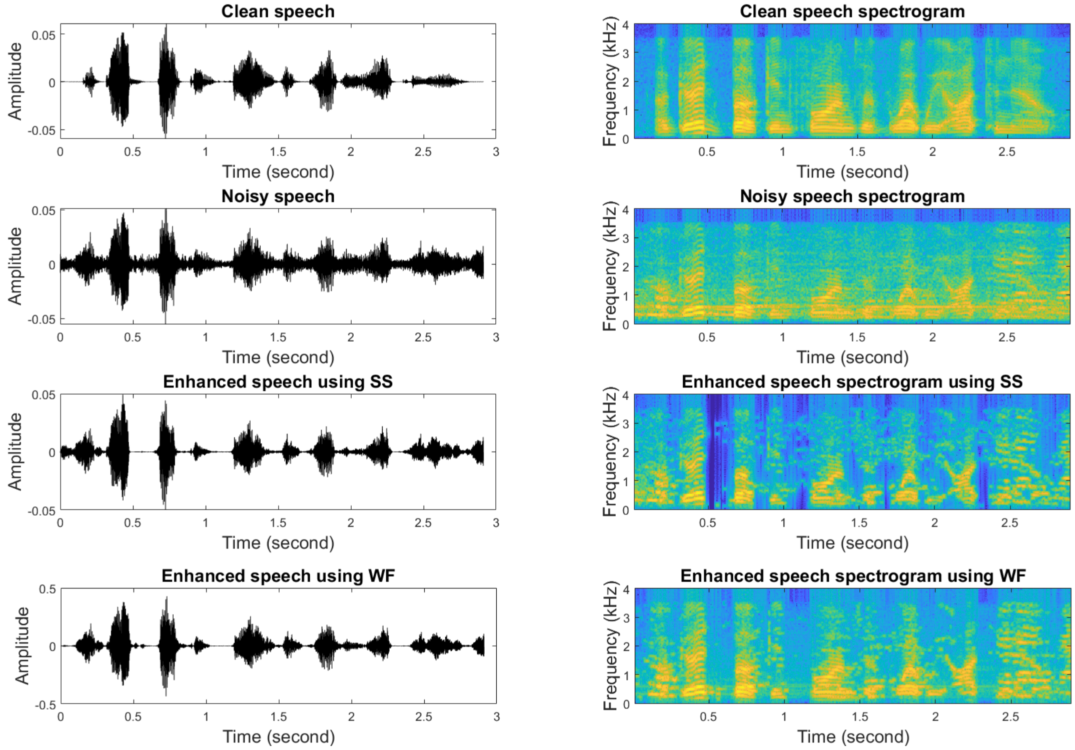

Fig. 3. Time domain and spectrogram representations of clean speech, noisy speech, enhanced speech using SS and enhanced speech using WF for the speech uttered by a male speaker and degraded by the airport noise at 5dB.

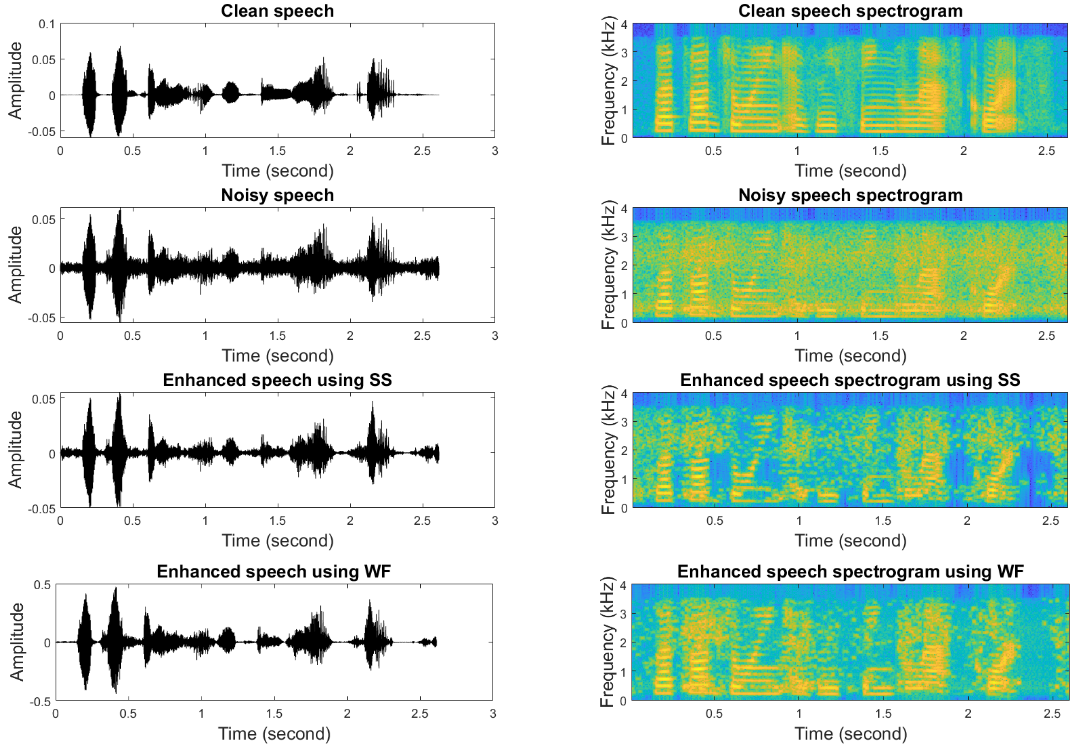

Fig. 4. Time domain and spectrogram representations of clean speech, noisy speech, enhanced speech using SS and enhanced speech using WF for the speech uttered by a female speaker and degraded by the train noise at 5dB.

[8] S. S. Bharti, M. Gupta, and S. Agarwal, “A new spectral subtraction method for speech enhancement using adaptive noise estimation,” in *3rd IEEE International Conference on Recent Advances in Information technology (RAIT)*, 2016, pp. 128–132.

[9] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall PTR, 1993.

[10] M. A. Abd El-Fattah, M. I. Dessouky, A. M. Abbas, S. M. Diab, E.-S. M. El-Rabaie, W. Al-Nuaimy, S. A. Alshebeili, and F. E. Abd El-Samie, “Speech enhancement with an adaptive Wiener filter,” *International Journal of Speech Technology*, vol. 17, no. 1, pp. 53–64, 2014.

[11] K. Khaldi and H. Touati, “Speech enhancement in EMD domain using spectral subtraction and Wiener filter,” *5th International Conference on Control Engineering and Inf. Technology*, vol. 32, pp. 27–32, 2018.

[12] P. Lei, M. Chen, and J. Wang, “Speech enhancement for in-vehicle voice control systems using wavelet analysis and blind source separation,” *IET Intelligent Transport Systems*, vol. 13, no. 4, pp. 693–702, 2018.

[13] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[14] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2013.

[15] D. Comminiello and J. C. Príncipe, *Adaptive learning methods for nonlinear system modeling*. Butterworth-Heinemann, 2018.

[16] S. W. Smith, *The scientist and engineer’s guide to digital signal processing*. California Technical Publishing, 1999, vol. 14.

[17] E. W. Healy, M. Delfarah, J. L. Vasko, B. L. Carter, and D. Wang, “An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker,” *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4230–4239, 2017.

[18] Y. Hu and P. C. Loizou, “Subjective comparison of speech enhancement algorithms,” in *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 1, 2006, pp. 153–156.

[19] “ITU-T Recommendation P.800: Methods for subjective determination of transmission quality,” 1996.