Circuits, Systems, and Signal Processing (2024) 43:2891–2908
https://doi.org/10.1007/s00034-023-02549-2

Optimized Sigmoid Functions for Speech Presence Probability and Gain Function in Speech Enhancement

Hai Huyen Dam[1] · Sven Nordholm[1] · Pei Chee Yong[2] · Siow Yong Low[3]

Received: 23 November 2022 / Revised: 18 October 2023 / Accepted: 21 October 2023 /
Published online: 22 January 2024

Abstract

Speech presence probability (SPP) and gain functions such as Wiener filter or MMSE estimators require an estimate of the a-priori signal-to-noise ratio (SNR). However, the estimation of the a-priori SNR is computationally involved and sensitive to noise variations. This paper proposes to approximate the SPP and the overall gain function of a speech enhancement system by using sigmoid functions to reduce the need of estimating the a-prior SNR. By applying an approximation via the sigmoid functions it is shown that only the a-posteriori estimate of SNR is needed, resulting in a low complexity system. The sigmoid function is designed with an optimization algorithm to optimize its parameters with respect to speech quality measures. The optimization algorithm is based on the idea that the solution obtained for a given problem should move towards the best solution and avoid the worst solution. The proposed algorithm requires minimal control parameters and does not require any algorithm specific parameters. Simulation results show that the proposed sigmoid functions achieve good

Sven Nordholm, Pei Chee Yong and Siow Yong Low have been contributed equally to this work.

✉ Hai Huyen Dam
H.Dam@curtin.edu.au

Sven Nordholm
S.Nordholm@curtin.edu.au

Pei Chee Yong
peichee.yong@nuheara.com

Siow Yong Low
sy.low@soton.ac.uk

[1] School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Kent Street, Bentley, Perth, WA 6102, Australia

[2] Nuheara, Perth, WA, Australia

[3] Connected Intelligence Research Group (CIRG), University of Southampton Malaysia, 79100 Iskandar Puteri, Johor, Malaysia

results in terms of speech quality measures when compared with existing methods while providing significantly lower complexity for implementation.

Keywords Single channel speech enhancement · A-priori SNR estimation · Decision directed approach · Optimization

1 Introduction

The prevalence of smart devices in our daily lives has pushed for an unprecedented demand on audio communication systems. As such, the need for a seamless speech communication system on such devices especially in noisy environments is highly sought after. An effective way to enhance noisy speech is via single channel speech enhancement techniques [1, 6, 7]. From the ideas of spectral subtraction by Boll [1], more optimal methods were developed that optimize MMSE and log MMSE errors [6, 7]. Those methods have highlighted the two main tasks associated with single channel processing which are noise suppression and speech preservation. However, it is a challenge to achieve both tasks optimally as suppression and distortion are conflicting measures, which results in a natural trade-off [13, 14, 19, 20]. For instance, if the noise estimator makes an erroneous estimation in the noise statistics, it will cause a mismatch in the noise suppression function. This in turn generates annoying musical artefacts, which reduce the overall perceptual quality of the enhanced speech [1, 12, 18]. An efficient way to combat the musical noise problem is to improve the noise spectrum estimation [3, 8, 12]. By using a soft voice activity detector idea based on the speech presence probability (SPP), significant improvement of the noise spectrum estimation was achieved. Yong et al. [22] further improved upon those results by using a modified sigmoid function which incorporates an a-priori SNR estimate to reduce the latency of the real-time SNR estimation. The modified decision directed approach [23] overcomes the one-frame delay problem when estimating the a-priori SNR by matching the estimated clean speech spectrum with the *a-priori* SNR as opposed to the previous frame. The reduction in SNR estimation's latency results in greater noise suppression and generates less musical noise.

While [22, 23] outlined a means to improve the SNR estimation, the method still employed the *a-priori* SNR estimate which was computationally complex and often gave large variations in the estimate for non-stationary background noise. Enzner [4] addressed the *a-priori* SNR problem by using a Bayesian Marginalization technique, but this required a lot of pre-training. In addition, it required the estimation of the global *a-priori* SNR of the speech data for each SNR. The result from this is a look up table that can be related to the *posteriori* SNR. Enzner did not device a way to address the noise estimation problem.

In this paper, we propose to overcome the aforementioned problems by using the modified sigmoid function to approximate both the speech presence probability (SPP) and the speech enhancement gain function and illustrated through the Wiener filter. The benefit is twofold. First, by applying an approximation via the sigmoid functions it is shown that only the *posteriori* estimate of SNR is needed, resulting in a lower complexity system which does not require the a-priori SNR estimation. Secondly, since only the *posteriori* information is needed, the proposed method can directly measure the variations in non-stationary noise scenarios, thereby reducing its sensitivity to large variations typically observed in non-stationary noise. The sigmoid function is designed with four parameters to be optimized. This paper further employs an efficient optimization algorithm, which optimizes the parameters of the sigmoid function with respect to speech quality measures. By incorporating speech quality measures in the optimization, the set of optimized parameters yield the best possible perceptually enhanced and intelligible speech. The optimization algorithm is based on the idea that the solution obtained for a given problem should move towards the best solution and avoid the worst solution. The proposed algorithm requires minimal control parameters and does not require any algorithm specific parameters.

Simulation results shows the comparison in performance of several speech quality measures, namely the perceptual evaluation of speech quality (PESQ) measure [17], the short-time objective intelligibility (STOI) measure [21] and the log-likelihood ratio (LLR) [15] for (i) the decision directed, (ii) modified decision directed, (iii) and the system with the sigmoid functions for both the gain function and the speech present probability. The proposed method is tested on some of the common types of noise, namely the babble, factory, pink and white noise. The set of sigmoid function coefficients are optimized with 0 dB SNR and the system is tested for various SNR. The results demonstrate that the proposed sigmoid functions achieve better results in terms of PESQ, STOI and LLR when compared with existing methods, namely the decision directed and modified decision directed with low complexity. In addition, a trade-off between PESQ, STOI and LLR performance can be achieved between the two proposed optimized sigmoid gain functions.

The paper is organized as follows: The system mode and the gain function are discussed in Sect. 2. The a-priori SNR estimation and the speech present probability are investigated in Sect. 3. The proposed system with the sigmoid function model for both the SPP and the gain function is given in Sect. 4. The optimization procedure is given in Sect. 5. Simulation results are given in Sect. 6, and finally, the conclusions are in Sect. 7.

2 System Model and the Gain Function

The goal of speech enhancement scheme is to estimate the enhanced speech signal $\hat{x}(n)$, given a noisy signal $y(n) = x(n) + v(n)$, where $x(n)$ and $v(n)$ denote the clean speech signal and the noise, respectively. By applying the short-time Fourier transform (STFT) to the time data, the STFT of the noisy signal is given as

$$Y(k, m) = X(k, m) + V(k, m) \tag{1}$$

where $X(k, m)$ and $V(k, m)$ denote the STFT of the clean speech signal $x(n)$ and the uncorrelated additive noise $v(n)$, respectively [6, 7]. Here, k is the frequency bin index and m is the frame index. The estimated clean speech spectrum $\hat{X}(k, m)$ is then obtained as

$$\hat{X}(k,m) = G(k,m)Y(k,m) \tag{2}$$

where $G(k,m)$ is a spectral gain function. Our objective is to obtain an efficient and low complexity method to estimate the gain function $G(k,m)$. In the following, we discuss the methods to estimate $G(k,m)$.

The gain function $G(k,m)$ is often derived from MMSE or Log-MMSE optimization criteria [6], [7], which requires the estimation of the a-priori SNR. One popular MMSE method results in the Wiener filter [19], where the again function can be computed as

$$G_{\mathrm{WF}}(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)} \tag{3}$$

and $\xi(k,m)$ is the *a-priori* SNR, obtained as

$$\xi(k,m) = \frac{\lambda_x(k,m)}{\lambda_v(k,m)}. \tag{4}$$

Here, $\lambda_x(k,m)$ and $\lambda_v(k,m)$ represent the clean speech power spectral density and the noise power spectral density, respectively, which are unknown in practice and hence required to be estimated.

The gain function derived using the Log-MMSE criteria also requires the estimation of the a-priori SNR [24]. In [2], [24], the sigmoid function was investigated as the function of the a-priori SNR to model the gain function $G(k,m)$. However, the estimation of the a-priori SNR is often computationally complex [2]. In the following, we will discuss the estimation of the a-priori SNR and the speech presence probability that is used to estimate the noise power spectral density in (4).

3 A-priori SNR Estimation and the Speech Presence Probability

In [5], the a-priori SNR is estimated using the decision direction (DD) method,

$$\hat{\xi}_{DD}(k,m) = \max\left\{\beta\frac{|\hat{X}(k,m-1)|^2}{\hat{\lambda_v}(k,m)} + (1-\beta)P[\gamma(k,m)-1], \epsilon_0\right\} \tag{5}$$

where $\hat{X}(k,m-1)$ and $\hat{\lambda_v}(k,m)$ denote, the estimated clean speech spectrum and the estimated noise PSD, respectively. In addition, the parameter β denotes the smoothing factor, $P[\cdot]$ denotes the half-wave rectification and ϵ_o is the SNR floor. Here, $\gamma(k,m)$ is the *a-posteriori* SNR obtained as

$$\gamma(k,m) = \frac{|Y(k,m)|^2}{\lambda_v\,(k,m)}.$$

The modified decision direction method (MDD) was developed in [24] for the estimation of the a-priori SNR to improve further the speech quality of the DD method.

The main difference between the MDD and DD methods is the estimation of the a-priori SNR which requires the use of the gain function $G(k, m-1)$ in the previous iteration to estimate $\hat{X}(k, m-1)$,

$$\hat{\xi}_{MDD}(k,m) = \max\left\{\beta\frac{|G(k,m-1)Y(k,m)|^2}{\hat{\lambda}_v(k,m)} + (1-\beta)P[\gamma(k,m)-1], \epsilon_0\right\}. \quad (6)$$

In addition, the estimations in (5) and (6) require the estimation of the a-posteriori SNR and the noise power spectral density $\lambda_v(k,m)$. One common method of estimating $\lambda_v(k,m)$ is applying a temporal recursive smoothing to the noisy observation using the speech presence probability (SPP) $p(k,m)$ [3], [11],

$$\lambda_v(k,m) = p(k,m)\lambda_v(k,m-1) + (1-p(k,m))\,||Y(k,m)||^2. \quad (7)$$

Assuming both $X(k,m)$ and $V(k,m)$ have Gaussian distributions, then the SPP is given by [3],

$$p(k,m) = \left\{1 + (1+\xi(k,m))\,Q\exp\left(-\gamma(k,m)\frac{\xi(k,m)}{1+\xi(k,m)}\right)\right\}^{-1} \quad (8)$$

where $Q = \dfrac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}$ is the ratio between $P(\mathcal{H}_0)$ the *a-priori* probability for speech absence and $P(\mathcal{H}_1)$ the probability for speech presence.

It can be seen that the speech presence probability $p(k,m)$ yields a value that is close to one when $\gamma(k,m)$ is sufficiently large and is small otherwise. In between zero and one, a soft transition for SPP is desired. As such, a sigmoid function is employed for the SPP in Eq. (8) [23] as a function of the estimated a-posteriori SNR $\gamma(k,m)$ and fixed coefficients

$$p_{\text{sig}}(k,m) = \frac{1}{1+e^{-c_{\text{sig}}(\gamma(k.m)-d_{\text{sig}})}} \quad (9)$$

where c_{sig} and d_{sig} indicate, respectively, the slope and the mean of the sigmoid function, given by

$$c_{\text{sig}} = \frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}, \quad d_{\text{sig}} = \log\left(Q\left(1+\xi_{\mathcal{H}_1}\right)\right)\frac{1+\xi_{\mathcal{H}_1}}{\xi_{\mathcal{H}_1}}. \quad (10)$$

The value $\xi_{\mathcal{H}_1}$ denote the *a-priori* SNR when speech is present.

The estimations of the a-priori SNR $\xi(k,m)$ and the speech presence probability in (5), (6), (8) are computational expensive and can be sensitive to large variations in the noise estimate. From (8), it is evident that if only the a-posteriori SNR $\gamma(k,m)$ is used, it is easier to control the variations in the noise power. Thus, we propose to model $G(k,m)$ estimation based on the a-posteriori SNR $\gamma(k,m)$ for each frequency bin k and time instance m. This will result in a lower complexity estimator as only the noise and noisy speech are required to be estimated. In addition, a general sigmoid function is proposed for the SPP in (9) and the sigmoid function coefficient will be optimized to improve the performance.

4 The Proposed Gain Function and Speech Presence Probability

In this section, we propose to approximate the gain function and the speech presence probability as general sigmoid functions of the a-posteriori SNR $\gamma\,(k, m)$. The gain function for each frequency bin k and instant time m can be obtained as

$$G_{\mathrm{SIG}}(k, m) = \max\left\{\frac{2}{1 + e^{-a(\gamma(k,m)-b)}} - 1, 0\right\}. \tag{11}$$

where a and b are some constants. In addition, the speech presence probability can also be modelled as a general sigmoid function

$$p_{\mathrm{SIG}}(k, m) = \frac{1}{1 + e^{-c(\gamma(k,m)-d)}}. \tag{12}$$

where c and d are constant parameters, which can be optimized. This will result in lower complexity for the estimation as the a-posterior SNR $\gamma(k, m)$ is much easier to estimate.

It has been reported in [23] that if the SPP estimate $p(k, m)$ is used directly in Eq. (8), then the noise estimate becomes more noisy due to large variations in $p(k, m)$ which modulates the noise estimate. One way to reduce this variability is to smooth $\gamma(k, m)$ or $p(k, m)$. However, the smoothing results in extra delay, which reduces its noise tracking capability. Here, we quantize $p_{\mathrm{appr}}(k, m)$ into four different regions, i.e.,

$$p' = \begin{cases} \text{noise only presence}, \mathcal{P}_1, & p_{\mathrm{appr}} \leq p_1 \\ \text{likely speech presence}, \mathcal{P}_2, & p_1 < p_{\mathrm{appr}} \leq p_2 \\ \text{more likely speech presence}, \mathcal{P}_3, & p_2 < p_{\mathrm{appr}} \leq p_3 \\ \text{most likely speech presence}, \mathcal{P}_4, & p_{\mathrm{appr}} \geq p_3 \end{cases} \tag{13}$$

where $0 < p_1 < p_2 < p_3 \leq 1$ are different values of the sigmoid function, they correspond to an instantaneous estimate of the SPP. These quantized values are mapped to different averaging smoothing constant. For the region where speech is less likely to present, i.e. when $\gamma \approx 1$ (this means 0 dB), the averaging constant for the noise estimation should be fast. The result is an even smoothed estimate compared to the original noise PSD estimate when γ is small, which reduces the likelihood of noise being overestimated and underestimated locally. For the regions where speech is either more likely or most likely to present, the soft transitions of p_{appr} might not be sufficient for the noise PSD estimate to change from using the previous noise PSD estimates to tracking the current noisy observations and vice versa. Accordingly, to avoid those pitfalls, quantized decisions are imposed on p_{appr} to realize an improved *posterior* SPP estimate.

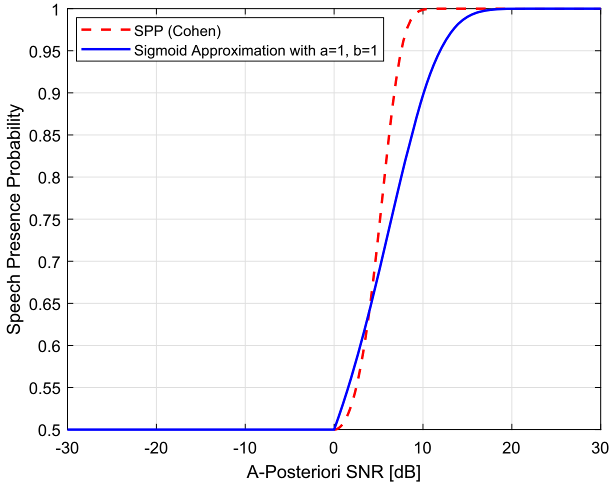

Fig. 1 Approximation of the Wiener filter using Eq. (11) with $a = 1$ and $b = 1$

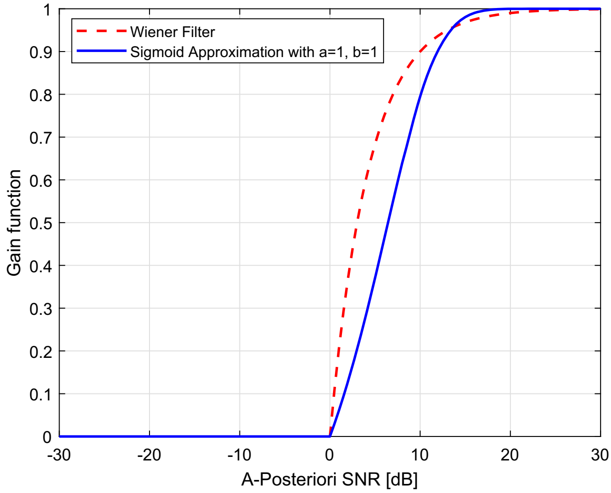

Fig. 2 Approximation of the speech presence probability $q = 0.5$ in Eq. (8) using Eq. (12) with $a = 1$ and $b = 1$

Now we have replaced the *a-priori* SNR with the *posterior* SNR through our approximation. How well the approximation works for the Wiener filter (3) and the SPP in Eq. (8) is shown in an example, by choosing the coefficients $\mathbf{x} = [a\ b\ c\ d] = [1\ 1\ 1\ 1]$ and $Q = 0.5$, see Figs. 1 and 2. It can be seen that the approximation is very close for the SPP and relatively close for the Wiener filter approximation with the sigmoid approximation being slightly more aggressive for $a = 1$.

However, the main benefit is that we can optimize these coefficients based on data which generalizes them to a more flexible data based functions. Hence, we investigate on how to optimize the coefficient vector for unknown, $\mathbf{x} = [a\ b\ c\ d]$. It is proposed that the optimization is made with respect to the maximum achievable speech quality measures as that will naturally provide the best objective evaluated enhanced speech. In general, the speech quality assessment can be classified in terms of subjective and objective measures. Subjective evaluation involves subjective listening test by some listeners while objective evaluation measures the numerical distance between the reference signal and the processed signal. One established method of evaluating the enhanced signal is using perceptual evaluation of speech quality (PESQ). PESQ is an automatic computation algorithm to replace human subjects in the evaluation of the mean opinion score (MOS). The PESQ model considers how human perceive speech and it has been widely used in the evaluation of speech quality. Another popular measure is the short-time objective intelligibility (STOI) measure, which highly correlates with the intelligibility of speech. By optimizing with respect to both PESQ and STOI, the parameters are optimized to give the speech an overall quality improvement and speech intelligibility. Thus, a multi-objective optimization problem can be formulated with PESQ and STOI as the objective measure,

$$\begin{cases} \max \quad f(\mathbf{x}) = \mathrm{PESQ}(\mathbf{x}) + \alpha \mathrm{STOI}(\mathbf{x}) \\ \text{subject to} \quad \mathbf{x}_l \leq \mathbf{x} \leq \mathbf{x}_u \end{cases} \tag{14}$$

where $\mathbf{x}_l$ and $\mathbf{x}_u$ are the lower and the upper bounds for the coefficient vector $\mathbf{x}$, respectively, and α is the weighting constant. Different value of α results in different optimal solution for the Pareto optimality, allowing the trade-off between the two objective measures.

5 Optimization Procedure

In this paper, the Jaya method [16] with a modified stopping criteria is employed to obtain the optimal solution to the optimization problem (14). At any iteration k, we have N number of candidate solutions. Let the best candidate obtain the best value of $f(\mathbf{x})$ and the worst candidate obtain the worse value of $f(\mathbf{x})$,

$$\begin{aligned} \mathbf{x}_{best,k} &= \arg\max_i f(\mathbf{x}_{k,i}) \\ \mathbf{x}_{worse,k} &= \arg\min_i f(\mathbf{x}_{k,i}). \end{aligned} \tag{15}$$

The coefficient vectors of the $k+1$ iteration are given as

$$\mathbf{x}_{k+1,i} = \mathbf{x}_{k,i} + r_{1,k,i}\left(\mathbf{x}_{best,k} - |\mathbf{x}_{k,i}|\right) + r_{2,k,i}\left(\mathbf{x}_{worst,k} - |\mathbf{x}_{k,i}|\right) \tag{16}$$

where $r_{1,k,i}$ and $r_{2,k,i}$ are random numbers in the range $[0, 1]$. The first term in Eq. (16) indicates the tendency for the solution to move closer to the best solution while the second term indicates the tendency to avoid the worst solution. $\mathbf{x}_{k+1,i}$ is accepted if it gets a better solution. The algorithm stops if the difference in the optimal objective function between the two consecutive iterations is small. The steps for the optimization algorithm are summarized in Procedure 1.

Procedure 1: Optimization algorithm

- Step 1: Initialize the coefficient vector $\mathbf{x}_{0,i},\ 1 \leq i \leq N$ for the 0^{th} iteration. Set $k = 0$.
- Step 2: Calculate the objective function $f(\mathbf{x}_{k,i})$. Obtain the best and the worse solutions $\mathbf{x}_{best,k}$ and $\mathbf{x}_{worst,k}$ as in (15).
- Step 3: Obtain the new set of coefficient vectors for the $k+1$ iteration as in (16). For all the value $1 \leq i \leq N$, if $f(\mathbf{x}_{k+1,i}) < f(\mathbf{x}_{k,i})$, then set $\mathbf{x}_{k+1,i} = \mathbf{x}_{k,i}$. Otherwise, $\mathbf{x}_{k+1,i})$ remains the same as before.
- Step 4: The algorithm converges if there is no improvement in the maximum objective function or the maximum number of iterations is reached. Otherwise, set $k := k + 1$ and return to Step 2.

6 Experimental Results

For the objective evaluation, the noisy speech corpus NOIZEUS with 30 IEEE speech sequences were employed [9, 10]. The database was chosen as it was developed to facilitate for algorithm comparison purpose. More information about the NOIZEUS can be found in [9]. The noisy speech was corrupted with babble, factory, pink and white noise for a wide range of SNRs. All the results are generated with $K = 256$ frequency bins with a sampling frequency of $f_s = 8000$. A square-root Hanning window was used with 50% overlap. Simulations are evaluated with

$$p' = \begin{cases} \mathcal{P}_1, & p_{\text{appr}} \leq 0.55 \\ \mathcal{P}_2, & 0.55 < p_{\text{appr}} \leq 0.7 \\ \mathcal{P}_3, & 0.7 < p_{\text{appr}} < 0.8 \\ \mathcal{P}_4 & p_{\text{appr}} > 0.8 \end{cases}$$

where $\mathcal{P}_i = \exp(-2.2R) / \left(t_i f_f\right)$ indicates the exponential smoothing constant, with $i = [1, 2, 3, 4]$. Here, R indicates the STFT frame rate, t_i denotes the averaging time constant, with $t_1 < t_2 < t_3 \ll t_4$. This means that the averaging time is mapped to the speech presence probability but the averaging times and thresholds can be modified.

To evaluate the performance of the proposed sigmoid gain function and proposed SPP, the problem (14) is optimized for the different type of noise, namely the babble, factory, pink and white noise, with signal-to-noise ratio of 0 dB. For each type of noise, the optimal set of coefficient is then tested for different levels of SNR. The SNR level is increased from −5 dB to 10 dB. The results are compared with those obtained from the decision directed method and the modified decision directed method. As mentioned earlier, the proposed method has significantly lower complexity than both the decision directed and modified decision directed methods as it does not require the estimation of the a-priori SNR.

Table 1 PESQ, STOI and LLR performance for different SNR with babble noise and $K = 256$

SNR	Methods	PESQ	STOI	LLR
−5 dB	Decision direct	1.4871	0.5321	1.3951
	Modified decision direct	1.4987	0.5427	1.3803
	Optimized with $\alpha = 0$	1.7029	0.5896	1.1551
0 dB	Decision direct	1.8497	0.6533	1.2074
	Modified decision direct	1.8589	0.6711	1.1919
	Optimized with $\alpha = 0$	2.0042	0.7051	0.9884
5 dB	Decision direct	2.2566	0.7675	0.9981
	Modified decision direct	2.2496	0.7881	0.8824
	Optimized with $\alpha = 0$	2.3815	0.8033	0.8106
10 dB	Decision direct	2.6527	0.8567	0.7964
	Modified decision direct	2.6461	0.8760	0.7451
	Optimized with $\alpha = 0$	2.7111	0.8705	0.6401

6.1 Performance Comparison Between the Proposed Method, The Decision Direct Method and the Modified Decision Directed Method

Table 1 shows the PESQ, STOI and LLR results for different speech enhancement methods: (i) the decision directed; (ii) the modified decision directed [22] and (iii) the result with the optimized gain function G_{SIG} and the weighting constant $\alpha = 0$. The coefficients for the gain function G_{SIG} are optimized with SNR= 0 dB and the results are tests for different SNR levels and babble noise. It can be seen from the table that the modified decision direct improves the PESQ, STOI and LLR results over the decision directed method. In addition, the optimized sigmoid gain function together with the sigmoid SPP improves the PESQ, STOI and LLR values further over the modified decision directed method. For example, at −5 dB SNR level, the optimized method with gain function G_{SIG} improves 0.2158 dB for PESQ over the decision directed method and 0.2041 dB over the modified decision directed method. For STOI measure, the optimized method improves 0.0575 dB and 0.0469 dB, respectively, over the decision directed and the modified decision directed methods. For the LLR measure, the optimized method is 0.24 dB and 0.168 dB lower than the decision directed and the modified decision directed methods, which means that the optimized method performs better than the other two methods. For other SNRs, the optimized method with sigmoid gain functions G_{SIG} also has significant improvement for PESQ, STOI and LLR over the decision directed and modified decision directed methods.

Tables 2, 3 and 4 show the results for the factory noise, pink noise and white noise for different SNR and different gain function methods. It can be seen that the optimized gain functions G_{SIG} have significant improvement for PESQ, STOI and LLR over the results obtained using the decision directed and the modified decision directed methods. For example, with SNR=−5 dB and white noise, the optimized method with the gain function G_{SIG} improves 0.2528 dB and 0.1643 dB for PESQ, respectively, over the decision direct method and the modified decision directed method. For the

Table 2 PESQ, STOI and LLR performance for different SNR with factory noise and $K = 256$

SNR	Methods	PESQ	STOI	LLR
−5 dB	Decision direct	1.6775	0.5527	1.1615
	Modified decision direct	1.6861	0.5574	1.1726
	Optimized with $\alpha = 0$	1.7343	0.5814	1.1501
0 dB	Decision direct	2.0850	0.6597	1.0122
	Modified decision direct	2.0967	0.6764	1.0108
	Optimized with $\alpha = 0$	2.1412	0.7061	0.9627
5 dB	Decision direct	2.4651	0.7644	0.8594
	Modified decision direct	2.4886	0.7888	0.8485
	Optimized with $\alpha = 0$	2.5202	0.8085	0.7902
10 dB	Decision direct	2.8375	0.8547	0.7180
	Modified decision direct	2.8672	0.8761	0.6902
	Optimized with $\alpha = 0$	2.8684	0.8872	0.6098

Table 3 PESQ, STOI and LLR performance for different SNR with pink noise and $K = 256$

SNR	Methods	PESQ	STOI	LLR
−5 dB	Decision direct	1.7107	0.6013	1.1380
	Modified decision direct	1.7685	0.6087	1.1351
	Optimized with $\alpha = 0$	1.8725	0.6335	1.1447
0 dB	Decision direct	2.1305	0.6940	0.9879
	Modified decision direct	2.1777	0.7125	0.9775
	Optimized with $\alpha = 0$	2.2801	0.7512	0.9528
5 dB	Decision direct	2.5036	0.8156	0.8251
	Modified decision direct	2.5585	0.8156	0.8251
	Optimized with $\alpha = 0$	2.6354	0.8451	0.7749
10 dB	Decision direct	2.8600	0.8662	0.7415
	Modified decision direct	2.9167	0.8894	0.7058
	Optimized with $\alpha = 0$	2.9394	0.9033	0.6544

STOI measure, the optimized method improves 0.0272 dB and 0.0204 dB, respectively, over the decision directed and the modified decision directed methods. For the LLR measure, the optimized method improves 0.09 dB over the decision directed and the modified decision directed methods. For all the cases, the optimization algorithm converges quickly which requires only a few iterations for convergence.

Figures 3, 4 and 5 show PESQ, STOI and LLR values for different speech enhancement methods with the babble noise and different SNRs. It can be seen that proposed method with the gain function G_{SIG} improve the results over the decision directed and the modified decision directed methods.

Table 4 PESQ, STOI and LLR performance for different SNR with white noise and $K = 256$

SNR	Methods	PESQ	STOI	LLR
−5 dB	Decision direct	1.4704	0.6044	1.4908
	Modified decision direct	1.5589	0.6112	1.4858
	Optimized with $\alpha = 0$	1.7232	0.6316	1.3973
0 dB	Decision direct	1.9910	0.6969	1.2862
	Modified decision direct	2.0552	0.7128	1.2789
	Optimized with $\alpha = 0$	2.1253	0.7331	1.2166
5 dB	Decision direct	2.3724	0.7833	1.1388
	Modified decision direct	2.4240	0.8016	1.1295
	Optimized with $\alpha = 0$	2.4772	0.8166	1.0483
10 dB	Decision direct	2.6890	0.8524	1.0142
	Modified decision direct	2.7428	0.8704	0.9869
	Optimized with $\alpha = 0$	2.7832	0.8821	0.8670

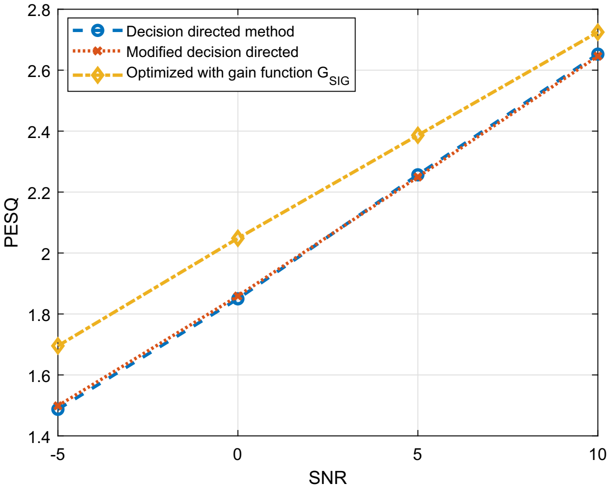

Fig. 3 PESQ for different speech enhancement methods with babble noise and different SNR for $\alpha = 0$

6.2 Trade-Off Investigation Between Perceptual Measures PESQ, STOI and LLR for Different Weighting Constants α and Different SNR

We now investigate the Pareto trade-off for different weighting factor α on the perceptual measures PESQ, STOI and LLR. Table 5 shows the trade-off between PESQ, STOI and LLR values for different weighting constraint α and the babble noise. The SNR level increases from −5 dB to 10 dB and the weighting constant α increases from 0 to 15. It can be seen from the table that there is a trade-off between the PESQ and STOI values. The PESQ values decrease when α increases while the STOI values increase. This is to be expected as the weighting provides an engineering choice between quality

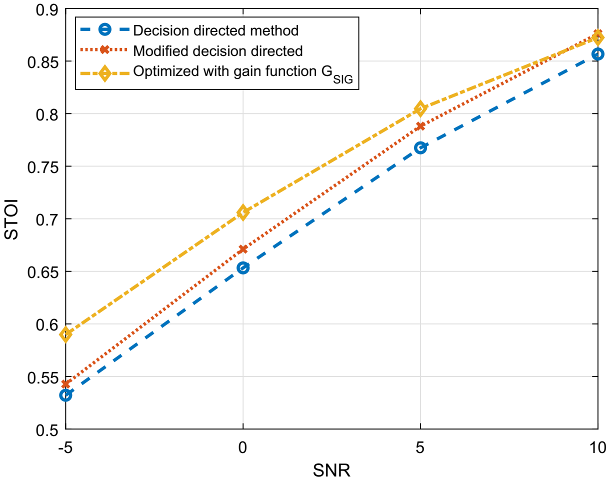

Fig. 4 STOI for different speech enhancement methods with babble noise and different SNR for $\alpha = 0$

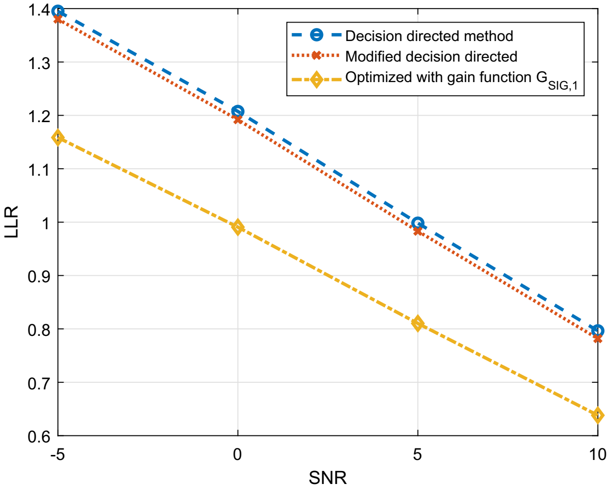

Fig. 5 LLR for different speech enhancement methods with babble noise and different SNR for $\alpha = 0$

and intelligibility through the PESQ and STOI measures, respectively. The LLR values are approximately the same for all the cases with the babble noise. When compared to the decision directed and modified decision directed performance in Table 1, the optimized sigmoid gain function has better PESQ, STOI and LLR performance than the decision directed method and the modified decision directed method.

Tables 6, 7 and 8 show the PESQ, STOI and LLR results for different SNRs and different weighting constant α with factory noise, pink noise and white noise, respectively. Similar to the case with the babble noise, when α increases, the PESQ value decreases while the STOI value increases. It can be seen that the weighting α provides a trade-off between PESQ and STOI in the objective measures [see Eq. (14)]. As α

Table 5 Trade-off between PESQ, STOI and LLR for different weighting function α with different SNR and babble noise

SNR	Weighting constant α	PESQ	STOI	LLR
−5 dB	Weighting $\alpha = 0$	1.7029	0.5896	1.1551
	Weighting $\alpha = 5$	1.6758	0.5921	1.1681
	Weighting $\alpha = 10$	1.6438	0.5943	1.1693
	Weighting $\alpha = 15$	1.6282	0.5957	1.1646
0 dB	Weighting $\alpha = 0$	2.0042	0.7051	0.9884
	Weighting $\alpha = 5$	2.0390	0.7101	0.9940
	Weighting $\alpha = 10$	2.0073	0.7138	0.9886
	Weighting $\alpha = 15$	1.9901	0.7154	0.9828
5 dB	Weighting $\alpha = 0$	2.3815	0.8033	0.8106
	Weighting $\alpha = 5$	2.3865	0.8108	0.8025
	Weighting $\alpha = 10$	2.3631	0.8167	0.7894
	Weighting $\alpha = 15$	2.3471	0.8187	0.7822
10 dB	Weighting $\alpha = 0$	2.7111	0.8705	0.6401
	Weighting $\alpha = 5$	2.7424	0.8807	0.6193
	Weighting $\alpha = 10$	2.7278	0.8882	0.5999
	Weighting $\alpha = 15$	2.7127	0.8906	0.5909

The sigmoid coefficients are optimized with SNR=0 and then the performance is tested for different SNR

Table 6 Trade-off between PESQ, STOI and LLR for different weighting function α with different SNR and factory noise

SNR	Weighting constant α	PESQ	STOI	LLR
−5 dB	Weighting $\alpha = 0$	1.7343	0.5814	1.1501
	Weighting $\alpha = 5$	1.7273	0.5915	1.1496
	Weighting $\alpha = 10$	1.7203	0.5946	1.1383
	Weighting $\alpha = 15$	1.6872	0.5959	1.1515
0 dB	Weighting $\alpha = 0$	2.1412	0.7061	0.9627
	Weighting $\alpha = 5$	2.1173	0.7162	0.9534
	Weighting $\alpha = 10$	2.1039	0.7185	0.9393
	Weighting $\alpha = 15$	2.0739	0.7213	0.9492
5 dB	Weighting $\alpha = 0$	2.5202	0.8085	0.7902
	Weighting $\alpha = 5$	2.5111	0.8176	0.7660
	Weighting $\alpha = 10$	2.4981	0.8199	0.7489
	Weighting $\alpha = 15$	2.4606	0.8237	0.7497
10 dB	Weighting $\alpha = 0$	2.8684	0.8872	0.6098
	Weighting $\alpha = 5$	2.8604	0.8790	0.6589
	Weighting $\alpha = 10$	2.8605	0.8892	0.5872
	Weighting $\alpha = 15$	2.8251	0.8937	0.5819

The sigmoid coefficients are optimized with SNR=0 and then the performance is tested for different SNR

Table 7 Trade-off between PESQ, STOI and LLR for different weighting function α with different SNR and pink noise

SNR	Weighting constant α	PESQ	STOI	LLR
−5 dB	Weighting $\alpha = 0$	1.8725	0.6335	1.1447
	Weighting $\alpha = 5$	1.8292	0.6385	1.1520
	Weighting $\alpha = 10$	1.8085	0.6436	1.1512
	Weighting $\alpha = 15$	1.7845	0.6445	1.1525
0 dB	Weighting $\alpha = 0$	2.2801	0.7512	0.9528
	Weighting $\alpha = 5$	2.2614	0.7600	0.9418
	Weighting $\alpha = 10$	2.2449	0.7622	0.9395
	Weighting $\alpha = 15$	2.2170	0.7648	0.9367
10 dB	Weighting $\alpha = 0$	2.6354	0.8451	0.7749
	Weighting $\alpha = 5$	2.6430	0.8521	0.7598
	Weighting $\alpha = 10$	2.6304	0.8531	0.7524
	Weighting $\alpha = 15$	2.6066	0.8563	0.7437
15 dB	Weighting $\alpha = 0$	2.9394	0.9033	0.6544
	Weighting $\alpha = 5$	2.9632	0.9101	0.6226
	Weighting $\alpha = 10$	2.9677	0.9107	0.6075
	Weighting $\alpha = 15$	2.9553	0.9136	0.5924

The sigmoid coefficients are optimized with SNR=0 and then the performance is tested for different SNR

Table 8 Trade-off between PESQ, STOI and LLR for different weighting function α with different SNR and white noise

SNR	Weighting constant α	PESQ	STOI	LLR
−5 dB	Weighting $\alpha = 0$	1.7232	0.6316	1.3973
	Weighting $\alpha = 5$	1.6893	0.6389	1.4100
	Weighting $\alpha = 10$	1.6396	0.6459	1.4321
	Weighting $\alpha = 15$	1.6178	0.6488	1.4370
0 dB	Weighting $\alpha = 0$	2.1253	0.7331	1.2166
	Weighting $\alpha = 5$	2.1086	0.7382	1.2154
	Weighting $\alpha = 10$	2.0544	0.7448	1.2221
	Weighting $\alpha = 15$	2.0127	0.7477	1.2295
5 dB	Weighting $\alpha = 0$	2.4772	0.8166	1.0483
	Weighting $\alpha = 5$	2.4754	0.8200	1.0387
	Weighting $\alpha = 10$	2.4370	0.8255	1.0181
	Weighting $\alpha = 15$	2.3927	0.8282	1.0180
10 dB	Weighting $\alpha = 0$	2.7832	0.8821	0.8670
	Weighting $\alpha = 5$	2.7902	0.8837	0.8587
	Weighting $\alpha = 10$	2.7780	0.8884	0.8239
	Weighting $\alpha = 15$	2.7382	0.8910	0.8122

The sigmoid coefficients are optimized with SNR=0 and then the performance is tested for different SNR

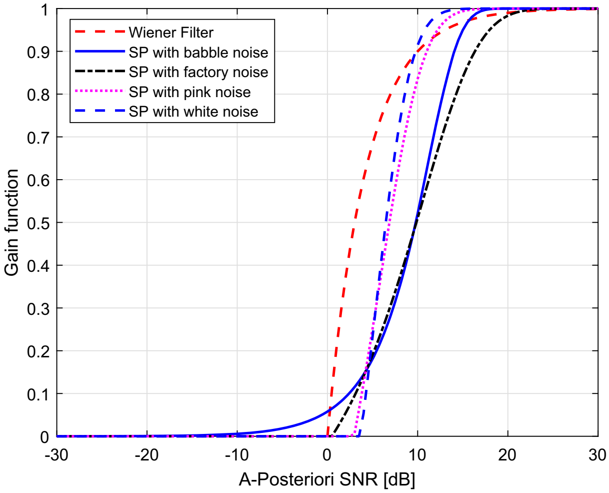

Fig. 6 Winner filter and optimized Sigmoid function for the gain function with different type of noise. The sigmoid functions are optimized from data at 0 dB

increases, more weighting is emphasized towards STOI as opposed to PESQ, which results in a higher value of STOI. The role of α serves as a trade-off between the two performance measures, which provides flexibility to the user to trade-off between the two measures. The increased of LLR in tandem with alpha shows that LLR is more correlated to STOI, which is related to the measure of speech intelligibility.

In addition, the LLR values improves slightly with a higher value of α. Similar to the babble noise case, the optimized sigmoid gain function achieves good trade-off performance when compared with the decision direct and modified decision directed methods. In addition, the proposed gain function has a lower complexity when compared with existing methods as it does not require the estimation of the a-priori SNR.

6.3 Approximation of the Gain Function and the Speech Present Probability using the Sigmoid Function for different Type of Noise

Figures 6 and 7 show the optimal sigmoid functions for the gain function and the speech present probability for different a-posteriori SNR. The optimal sigmoid functions for the gain function and the speech present probably are optimized together from data at 0 dB SNR for different type of noise, namely the babble noise, factory noise, pink and white noises. It can be seen from the figures that the optimized sigmoid function for the gain function and the speech present probability approximations follow the shape of the Wiener filter in Eq. (11) and the speech present probability in (8). In addition, the sigmoid functions for the factory and babble noises are slightly more aggressive than the sigmoid functions for the white and pink noises. The sigmoid functions are then tested for different SNR levels from −5 dB to 10 dB. It can be seen in Sects. 6.1 and 6.2 that the sigmoid models achieve good results for all the cases with a lower computational complexity as it does not require the estimation of the a-priori SNR.

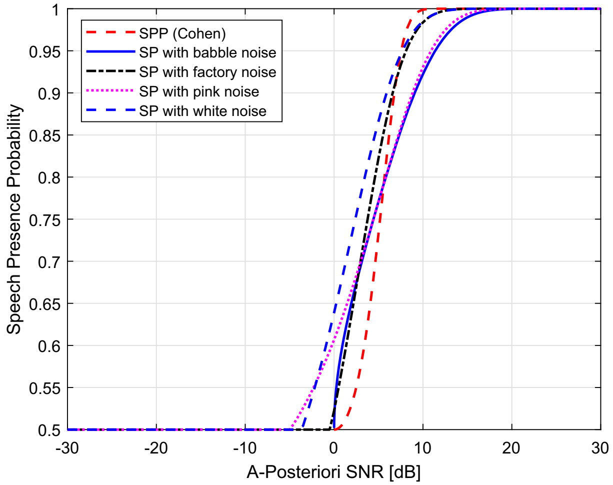

Fig. 7 Optimized sigmoid function for speech presence probability with different type of noise. The sigmoid functions are optimized from data at 0 dB

7 Conclusions

This paper proposes the use of sigmoid function for both the speech presence probability (SPP) and the overall gain function of a speech enhancement system as a means to achieve low complexity and efficient implementation. The former serves to better the SNR estimation and the latter provides an overall perceptually smooth gain function. The advantage of the proposed system is that it avoids the estimation the a-priori SNR resulting in an improved noise estimate. An efficient optimization algorithm is employed to solve the optimization problem, which optimizes the parameters of the sigmoid functions with respect to the speech quality measures. The optimization algorithm is based on the idea that the solution obtained for a given problem should move towards the best solution and avoid the worst solution. The presented algorithm requires minimal control parameters and does not require any algorithm specific parameters. Simulation results show that the proposed sigmoid functions achieve improved performance when compared with existing methods with low complexity.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

References

1. S. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**, 113–120 (1979)
2. K.Y. Chan, S. Nordholm, S.Y. Low, P.C. Yong, K.F.C. Yiu, A hybrid descent method for optimal sigmoid filter design. IEEE Signal Process. Lett. **21**(4), 478–482 (2014)
3. I. Cohen, Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003)
4. G. Enzner, P. Thune, Bayesian MMSE filtering of noisy speech by SNR marginalization with global PSD priors. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(12), 2289–2304 (2018)
5. Y. Ephraim, D. Malah, Speech enhancement using a minimummean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
6. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
7. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **33**(2), 443–445 (1985)
8. T. Gerkmann, R.C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking Delay. IEEE Trans. Audio Speech Language Process. **20**(4), 1383–1393 (2012)
9. Y. Hu, P. Loizou, Subjective evaluation and comparison of speech enhancement algorithms. Speech Commun. **49**, 588–601 (2007)
10. P. Loizou, *Speech Enhancement Theory and Practice* (CRC Press, Boca Raton, FL, 2007)
11. S. Y. Low, An insight into the rise time of exponential smoothing for speech enhancement methods, in *IEEE International Conference Signal Image Process Applications*, pp. 30–33 (2021)
12. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**, 504–512 (2001)
13. L. Nahma, P.C. Yong, H.H. Dam, S. Nordholm, An adaptive a-priori SNR estimator for perceptual speech enhancement. EURASIP J. Audio Speech Music Process. **1**, 1 (2019)
14. K. Paliwal, B. Schwerin, K. Wo, Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. Speech Commun. **54**(2), 282–305 (2012)
15. S. Quackenbush, T. Barnwell, M. Clements, *Objective Measures of Speech Quality* (Prientice Hall, Englewood Cliffs, 1988)
16. R.V. Rao, Jaya: a simple and new optimizaton algorithm for solving constrained and unconstrained optimization problems. Int. J. Eng. Comput. **7**, 19–34 (2016)
17. A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codec. IEEE Int. Conf. Acoust. Speech Signal Process. **2**, 749–752 (2001)
18. T. Rohdenburg, V. Hohmann, B. Kollmeier, Objective perceptual quality measures for the evaluation of noise reduction schemes, in *9th International Workshop on Acoustic Echo and Noise Control*, pp. 169–172 (2005)
19. P. Scalart, Speech enhancement based on a-priori signal to noise estimation, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96), 629–632 (1996)
20. M.K. Singh, S.Y. Low, S. Nordholm, Z. Zang, Bayesian noise estimation in the modulation domain. Speech Commun. **96**, 81–92 (2018)
21. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 125–2136 (2011)
22. P.C. Yong, S. Nordholm, H.H. Dam, Optimization and evaluation of sigmoid function with a priori SNR estimate. Speech Commun. **55**(2), 358–376 (2012)
23. P. C. Yong, S. Nordholm, H. H. Dam, Noise estimation based on soft decisions and conditional smoothing for speech enhancement, in *International Workshop on Acoustic Signal Enhancement* (2012)
24. P.C. Yong, S. Nordholm, H.H. Dam, Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement. Speech Commun. **55**(2), 358–376 (2013)