A hybrid descent method for optimal sigmoid filter design

Kit Yan Chan, Sven Nordholm, Siow Yong Low, Pei Chee Yong, Ka Fai Cedric Yiu

Abstract - In this letter, a hybrid descent method is used to determine a set of filter parameters for a sigmoid filter which attempts to work under various SNR conditions. It overcomes the limitations of the current sigmoid filters that performs effectively only at a single SNR. Results show that significant improvement in terms of better speech qualities can be achieved by the proposed sigmoid filter when working under various SNR conditions.

Index Terms - Noise suppression, sigmoid function, speech enhancement, hybrid optimization method, various SNR conditions

I. INTRODUCTION

Single-channel speech enhancement remains a viable choice for enhancing noisy speech due to its simplicity and convenience of implementation. Two main tasks associated with single-channel techniques are noise suppression and speech preservation. It remains a challenge to optimally achieve both tasks. It is made difficult when the noise characteristic is highly non-stationary [1]. In noisy environments, the noise estimator is prone to erroneous estimation in the noise statistics, thereby causing a mismatch in the suppression. The mismatch in the noise statistics also produces annoying musical artifacts, which reduce the intelligibility of the enhanced speech.

Recently, a speech enhancement scheme in the modulation domain has been proposed in order to reduce the musical noise [2]. As a result, a good trade-off between temporal slurring and musical noise can be tuned [2]. Also, progress has been made in the algorithmic development for SNR based cost functions and/or different prior statistical models to improve speech quality [3,5,7]. However, given their complexity, these aforementioned algorithms demand heavy computational requirements. This is critical for computationally-sensitive applications like cochlear implants and hearing aids.

Motivated by the desire to achieve low computational cost, Hu et al. [8] proposed the sigmoid (SIG) filter, which requires that various parameters be estimated. Hu et al. [8] showed that the SIG can be readily used to enhance the noisy signal as the SIG follows a pattern similar to the human listener’s psychometric function of intelligibility versus SNR. To further improve the effectiveness of the original SIG, Yong et al. [9] have developed a modified SIG (MSIG) with an a priori SNR estimate to cater real-time estimation of the SNR. It overcomes the one-frame delay when estimating the a priori SNR by matching the estimated clean speech spectrum and the a priori SNR as opposed to the previous frame. Results show that the modified SIG significantly outperforms the original a priori SNR based method and reduces audible noise and increases noise reduction for the commonly used gain functions, when the MSIG is developed under a particular SNR level.

Nevertheless, it is impractical to develop a single MSIG which can work effectively only at a single SNR level. It remains critical room for further improvement in determining the appropriate filter parameters. It is necessary to estimate appropriate filter parameters without having access to the statistics of SNR conditions. This paper proposes a hybrid descent algorithm (HAD) in order to determine appropriate filter parameters for various SNR conditions.

II. SIGMOID FILTER

Given a noisy signal \( y(t) = x(t) + v(t) \), with the clean speech, \( x(t) \), and the uncorrelated additive noise, \( v(t) \), the enhanced speech spectrum, \( \hat{X}(\omega, \ell) \), with respect to \( x(t) \) can be generated by the MSIG [9] namely \( G_{MSIG}(\omega, \ell, \kappa(\sigma)) \), as

\[
\hat{X}(\omega, \ell) = G_{MSIG}(\omega, \ell, \kappa(\sigma)) \cdot Y(\omega, \ell). 
\] (1)

In (1), \( Y(\omega, \ell) \) is the M-point short-time Fourier transformation of \( y(t) \); \( \omega \in [\omega_0, \omega_1, ..., \omega_{K-1}] \) is a real angular center frequency with \( K \) bands; \( \ell \in [0,1, ..., L-1] \) is the time frame index given by with \( L \) frames; and \( \kappa(\sigma) \) is the filter parameters used to control the characteristic of \( G_{MSIG} \). It works under the estimated SNR, namely \( \sigma \). \( G_{MSIG} \) is formulated as shown in (2):

\[
G_{MSIG}(\omega, \ell, \kappa(\sigma)) = \frac{1}{1 + e^{-k_1(\sigma)(\bar{\xi}(\omega, \ell) - s(\sigma))}} \cdot \frac{1 - e^{-k_2(\sigma)(\bar{\xi}(\omega, \ell))}}{1 + e^{-k_3(\sigma)(\bar{\xi}(\omega, \ell))}}, 
\] (2)

where \( \kappa(\sigma) \) is given as \( \kappa(\sigma) = [k_1(\sigma), k_2(\sigma), k_3(\sigma)] \); \( \bar{\xi}(\omega, \ell) \) is the a priori SNR estimated by the recently developed approach [9]. \( G_{MSIG} \) is effectively in real-time implementation when compared with the conventional decision-directed approach [5]; \( \sigma \) is estimated by the average of the a posteriori SNR which is given by:

\[
\sigma = E \left[ \frac{\|Y(\omega, \ell)\|^2}{E[\|Y(\omega, \ell)\|^2]} \right] = \frac{E[\|Y(\omega, \ell)\|^2]}{\lambda_c(\omega, \ell)}, 
\] (3)

with the M-point short-time Fourier transformation of the noise \( V(\omega, \ell) \) and the noise power spectral density \( \lambda_c(\omega, \ell) \). The first term of the \( G_{MSIG} \) represents a sigmoid function [8]. The slope and the mean of which are controlled by \( k_1(\sigma) \) and \( k_2(\sigma) \) respectively. They control the amount of musical noise, speech distortion and noise reduction of the enhanced speech. To balance the three quality measures, the sigmoid slope has to be sensitive towards speech and insensitive towards the variation of noise. The mean of the magnitude of the sigmoid function cannot be zero although \( \xi(\omega, \ell) \) is very small. Hence, poor noise reduction can be produced.
To improve the original SIG, the hyperbolic tangent function is introduced as the second term of $G_{\text{MSG}}$. Figure 1 shows the magnitude of $G_{\text{MSG}}$ with respect to $|\tilde{\zeta}_i(\omega, \ell)|$, where different magnitudes are obtained with different filter parameters. Hence, the magnitude is decreased when $|\tilde{\zeta}_i(\omega, \ell)|$ is low. Also, the magnitudes of $G_{\text{MSG}}$ are smaller than those of the original SIG, when $|\tilde{\zeta}_i(\omega, \ell)|$ is low. Hence, it provides more noise reduction at low $|\tilde{\zeta}_i(\omega, \ell)|$ than those of the original SIG.

![Figure 1 Characteristic of $G_{\text{MSG}}$ with different $|\tilde{\zeta}_i(\omega, \ell)|$](image)

The speech quality of the enhanced speech can be optimized with appropriate $\kappa(\sigma) = [k_1(\sigma), k_2(\sigma), k_3(\sigma)]$. Therefore, $\kappa(\sigma)$, is necessary to be optimized with respect to an specified speech quality measure, $\Theta(\ldots)$, by solving (4):

$$\text{Opt: } J_i(\kappa(\sigma)) = \Theta(\kappa(\sigma), X(\omega, \ell), Y(\omega, \ell)), \quad (4)$$

where $\Theta(\ldots)$ evaluates the speech quality of the enhanced speech $G_{\text{MSG}}(\omega, \ell, \kappa(\sigma))Y(\omega, \ell)$ with respect to the original speech $X(\omega, \ell)$. The speech quality can be evaluated by the objective measurement algorithms which are used to evaluate the performance of speech enhancement algorithms [12]. As perceptual evaluation of speech quality (PESQ), SNRseg, log-likelihood can be the commonly used measures, the optimization problem (4) can be either a minimization or maximization problem, depending on the speech quality measure $\Theta(\ldots)$.

However, the magnitude of $G_{\text{MSG}}$ varies with respect to $|\tilde{\zeta}_i(\omega, \ell)|$. As $|\tilde{\zeta}_i(\omega, \ell)|$ is related to the $a$ posteriori SNR, a particular filter parameter $\kappa(\sigma)$ may work satisfactorily only for a single $\sigma$. In this paper, the HDA is proposed in order to generate a set of filter parameters namely $\kappa = [\kappa_1(\sigma), \kappa_2(\sigma), \ldots, \kappa_N(\sigma)]$. Each $\kappa_i(\sigma)$ is determined with respect to a particular $\sigma$, with $i=1,2,\ldots N$. $\sigma_i \in [\sigma_1, \sigma_2, \ldots, \sigma_N] = \Omega$, where $\Omega$ is the set for the $a$ posteriori SNR and each $\sigma_i$ is calculated based on (3). The noisy speech, $Y(\omega, \ell)$, is contaminated artificially by adding $V(\omega, \ell)$ with $X(\omega, \ell)$ under an artificial SNR namely $\phi_i$, where $\phi_i < \phi_j \in [\phi_1, \phi_2, \ldots, \phi_N] = \Phi$ and $i < j$ with $i,j=1,2,\ldots N$, and all $\phi_i$ is within the predefined upper bound, $\phi_1$, and low bound, $\phi_N$.

### III. PARAMETER ESTIMATION FOR SIGMOID FILTERS

As the dimension of $J_i(\kappa(\sigma))$ is not high, the simplex search algorithm could be used to locate the optima [14]. However, the global optimum may not be located due to the nonlinearity of $J_i(\kappa(\sigma))$. To seek the global optimum, simulated annealing [15] could be used. As the convergence rate of simulated annealing is relatively slow, a HDA comprising the simulated annealing and the simplex search technique is proposed to solve $J_i(\kappa(\sigma))$. It is used to generate a set of appropriate $\kappa_i^{\text{opt}}(\sigma_i)$ with respect to $\sigma_i$. Each $\sigma_i$ is aligned with the artificial SNR $\phi_i$, where $X(\omega, \ell)$, is artificially corrupted by $V(\omega, \ell)$ with the artificial SNR, $\phi_i$.

In Step 1 of the HDA, $i$ is initialized by setting $i=1$. Step 2 creates $Y(\omega, \ell)$, artificially with the signal to noise ratio, $\phi_i$, and Step 3 estimates the $a$ posteriori SNR, $\sigma_i$. The HDA searches the filter parameter with respect to $\sigma_i$. It generates an initial guess, $\kappa_i(\sigma_i)$, randomly with $k=0$ and $i=1$. Then, Step 5 uses the simplex search method to determine the optimum for $\kappa_i(\sigma_i)$, and Step 6 uses the simulated annealing [10] to explore a solution, $\kappa_i(\sigma_i)$ which is better than the optimum $\kappa_i^{\text{opt}}(\sigma_i)$.

Steps 5 and 6 keep iterating mutually until no better solution can be found. The searching process with respect to $\sigma_i$ stops in Step 7, after converged. The converged solution $\kappa_i^{\text{opt}}(\sigma_i)$ is returned as the outcome for $\sigma_i$ with $\kappa_i^{\text{opt}}(\sigma_i) = \kappa_i(\sigma_i)$.

### Hybrid Descent Algorithm (HDA)

**Input:** $X(\omega, \ell), V(\omega, \ell), \Phi$ ; **Output:** $\kappa, \Omega$;

**Step 1:** Initialize by setting $i=1$ and $k=0$.

**Step 2:** Create $Y(\omega, \ell)$ by artificially adding $V(\omega, \ell)$ to $X(\omega, \ell)$ with the artificial SNR, $\phi_i$.

**Step 3:** Calculate the $a$ posteriori SNR, $\sigma_i$, based on (3).

**Step 4:** Generate $\kappa_i(\sigma_i)$ randomly within the operational range of the filter parameters and evaluate the enhanced speech quality based on $J_i(\kappa_i(\sigma_i))$.

**Step 5:** Search for the local optimum $\kappa_i^{\text{opt}}(\sigma_i)$ for $J_i(\kappa_i(\sigma_i))$ using the simplex search method with the initial guess $\kappa_i(\sigma_i)$, where $J_i(\kappa_i^{\text{opt}}(\sigma_i)) - J_i(\kappa_i(\sigma_i)) \leq -\epsilon_i$ with a positive $\epsilon_i$.

**Step 6:** Starting from $\kappa_i^{\text{opt}}(\sigma_i)$, execute the simulating annealing until a solution $\kappa_i(\sigma_i)$ is obtained, where $J_i(\kappa_i(\sigma_i)) - J_i(\kappa_i^{\text{opt}}(\sigma_i)) \leq -\delta_i$ with a positive parameter $\delta_i$.

**Step 7:** If $k$ is larger than a predefined value, set $\kappa_i^{\text{opt}}(\sigma_i) = \kappa_i(\sigma_i)$, and go to Step 8. Otherwise, set $k = k + 1$ and go to Step 5.
Step 8: If \( i < N \), then initialize with \( k = 0 \), put \( \kappa_i (\sigma_i) = \kappa^{opt} (\sigma_i) \), set \( i = i + 1 \) and go to Step 2.

Step 9: Create and return the two sets \( \Omega \) and \( K \) where
\[
\Omega = [\sigma_1, \sigma_2, ..., \sigma_N] \quad \text{and} \quad K = [\kappa^{opt} (\sigma_1), \kappa^{opt} (\sigma_2), ..., \kappa^{opt} (\sigma_N)].
\]

For \( \sigma_{i+1} \), the filter parameter, \( \kappa^{opt}(\sigma_{i+1}) \) is determined by performing Steps 3 to 8, while \( \kappa^{opt}(\sigma_i) \) is used as the initial guess. When \( i \) in Step 8 reaches \( N \), the sets for the filter parameters, \( \kappa^{opt}(\sigma_i) \in K \), and the \( a \ posteriori \) SNR, \( \sigma_i \in \Omega \), are created. \( \kappa^{opt}(\sigma_i) \) is used in \( G_{MSG} \) when the \( a \ posteriori \) SNR is closest to \( \sigma_i \).

IV. EXPERIMENTAL RESULTS

A. Experimental setup

The HDA was developed using the Matlab R2011b, whereby the routine ‘fincon’ is used to determine the local optimum of the filter parameters and the inbuilt parameters of the simulated annealing were selected based on [15] where the iteration number was large enough to avoid premature convergence.

The HDA was evaluated based on a database of noisy speech sequences, namely NOIZEUS [13], which were corrupted by pink noise, babble noise and factory noise in the ranges of -15 to 15 dB. Three commonly used speech quality measures, namely PESQ [12], SNRseg [16] and the log-likelihood (LLR) [16], were used to evaluate the performance of the enhanced speech.

The operational range for the SNR is defined between -15 dB and 15 dB. A set of filter parameters is determined with respect to the noisy signal contaminated with the SNR of -15dB, -14dB, … and 15dB respectively, where the filter parameters are optimized using the HDA. Using the optimized filter parameters, the filter is evaluated by the test set, where the test set was corrupted with SNR -14.5dB in steps of 1dB till 14.5dB. Here, the filter parameters which were optimized with respect to the nearest \( a \ priori \) SNR were used in the filter working under an untrained noisy environment.

B. Performance evaluation

Figure 2 shows the PESQ obtained by four \( G_{MSG} \) filters against the pink noise with respect to the test set: i) the adaptive \( G_{MSG} \), which uses the filter parameters optimized with respect to the training set; ii) the singly trained filter for low SNR namely \( G_{MSG}(-15) \) which was optimized with respect to -15dB; iii) another singly trained filter for high SNR namely \( G_{MSG}(15) \) which was optimized with respect to 15dB; and iv) the optimal \( G_{MSG} \), the filter parameters of which are optimized directly with respect to the test set. Hence, the performance of the adaptive \( G_{MSG} \) can be compared with the optimal one.

The results indicate that the PESQ obtained by the adaptive \( G_{MSG} \) are better than those obtained by the \( G_{MSG}(-15) \) and \( G_{MSG}(15) \), where \( G_{MSG}(-15) \) is close to the optimized PESQ with low SNRs and \( G_{MSG}(15) \) is close to those with high SNR. These results indicate that the singly trained filters, \( G_{MSG}(-15) \) and \( G_{MSG}(15) \), perform well only with respect to their own developed SNRs, while the adaptive \( G_{MSG} \) performs well for the whole range of SNRs. Hence, the adaptive \( G_{MSG} \) can align the optimized PESQ irrespective of the input SNR. Also, similar performance can be obtained by the adaptive \( G_{MSG} \) compared with the optimal \( G_{MSG} \).

Figures 3 and 4 show the SNRseg and LLR with respect to the pink noise. Similar results are produced when the adaptive \( G_{MSG} \) can obtain the best SNRseg and LLR which are closest to the optimized PESQ. Figures 5 to 7 show the speech quality measures for Factory noise, and Figures 8 to 10 show those for Babble noise. Similar results were obtained when the adaptive \( G_{MSG} \) can obtain the best speech qualities. These results indicate that the adaptive \( G_{MSG} \) performs well for the whole range of SNRs, while \( G_{MSG}(-15) \) and \( G_{MSG}(15) \) perform well only at their developed SNRs. Also, the performance of the adaptive \( G_{MSG} \) aligns with the optimal \( G_{MSG} \) under various SNR conditions.
V. CONCLUSION

A HDA is proposed to optimize the filter parameters of the MSIG, in order to enhance the quality of speech under various SNR conditions. The HDA generates a training set of filter parameters for some SNRs, and those parameters are used as the closest neighboring SNRs for the MSIG working in an untrained noisy environment. Results show that significant improvement in terms of the three speech quality measures can be achieved.

REFERENCES