A Novel Environment-Dependent Speech Enhancement Method with Optimized Memory Footprint

*Suhadi Suhadi, Sorel Stan, Tim Fingscheidt**

BenQ Mobile GmbH & Co. OHG, 81667 Munich, Germany
* Braunschweig Technical University, 38106 Braunschweig, Germany
suhadi.suhadi@benq.com, sorel.stan@benq.com, t.fingscheidt@tu-bs.de

10.21437/Interspeech.2006-69

Abstract

Data-driven speech enhancement (Fingscheidt and Suhadi [1]) aims at improving speech quality for voice calls in a *specific* noise environment. The essence of the method are a set of frequency-dependent weighting rules, indexed by *a priori* and *a posteriori* SNRs, which are learned from clean speech and background noise training data. The weighting rules must be stored for each frequency bin separately and take up about 400 kBytes memory, which makes DSP implementations relatively expensive.

In this paper we propose an alternative definition of the weighting rules which requires only 27 kBytes memory. That is 6.7% of the memory consumption of the original algorithm, with virtually no loss in performance measured in terms of speech distortion and noise attenuation. Our approach is to redefine the weighting rules on the Bark scale and store their parametric representation obtained by polynomial curve fitting.

Index Terms: speech enhancement, Bark scale, polynomial curve fitting.

1. Introduction

Environment noise can severely affect both speech quality and intelligibility for voice calls. The purpose of speech enhancement is to reduce the unwanted noise component as much as possible without introducing noticeable distortions of the useful speech signal.

The general approach to speech enhancement is to apply a weighting rule to the noisy speech spectral amplitudes for estimating the clean speech component. The derivation of the weighting rule is usually formulated as an optimization problem using error criteria such as Minimum Mean Square Error (MMSE) of spectral amplitudes, log-spectral amplitudes, or perceptually motivated variants of these [2, 3, 4, 5].

The spectra of clean speech and noise can be modeled using probability density functions (*pdf*). Ephraim and Malah [3, 4] have successfully employed Gaussian modeling of the real and imaginary part of the clean speech spectrum, and recent research shows that a Gamma *pdf* [6] or a Super-Gaussian *pdf* [7] lead to even better results.

The selection of the error criterion and of the *pdf* modeling the clean speech spectrum is essential, since wrong choices lead to higher residual noise and distortion of speech. To circumvent this problem, generalized estimators [1, 8] were derived to compute a weighting rule by considering *training* speech data instead of any explicit formulation of the clean speech spectrum *pdf*.

While Porter and Boll [8] derived the estimators under assumption of Gaussian-distributed noise, Fingscheidt and Suhadi [1] proposed a novel data-driven method for deriving the weighting rules, called Ideal Gain Averaging (IGA), which relies on *environment-specific training noise* instead of explicit modeling of noise spectral amplitude via a parametric *pdf*.

The weighting rules for each frequency bin are estimated separately during speech presence and absence using a voice activity detector (VAD), and stored in a look-up table indexed by the *a priori* and *a posteriori* SNRs [1]. However, the relatively high memory requirements of about 400 kBytes makes the implementation in embedded communication devices expensive and unattractive.

We propose a two-step approach for reducing the memory requirements of the weighting rules table: (1) the weighting rules are redefined on the Bark scale leading to Bark IGA (B-IGA), and (2) the new B-IGA weighting rules are parameterized using an R-order Polynomial Least Square (R-PLS) model. The combined approach leads to a new set of weighting rules which require only 6.7% compared to the original IGA algorithm, with virtually identical performance in terms of segmental Speech-to-Speech Distortion Ratio (SSDR) and segmental Noise Attenuation (NA) [1].

The remaining of our paper is organized as follows: Section 2 describes the derivation of the B-IGA weighting rules, followed by an outline of the R-PLS parameterization in Section 3. Section 4 presents the experimental results documenting the performance of the newly introduced algorithm, followed in Section 5 by the concluding remarks.

2. Bark-Scale Ideal Gain Averaging

2.1. Training the Weighting Rules

The noisy speech spectrum at frame l and frequency index k is explicitly computed in training by $Y_l(k) = X_l(k) + N_l(k)$, where $k = 0, \ldots, K-1$.

An estimate of the noise spectral variance $\hat{\lambda}_{N_l}(k)$ [9] together with the noisy speech spectrum $Y_l(k)$ are used to compute the *a priori* SNR $\hat{\xi}_l(k)$ using a modification of the decision-directed approach of Ephraim and Malah [1, 3]

$$\hat{\xi}'_l(k) = w\frac{|X_{l-1}(k)|^2}{\hat{\lambda}_{N_{l-1}(k)}} + (1-w)\max\left\{\hat{\gamma}_l(k) - 1, 0\right\},$$
$$\hat{\xi}_l(k) = \max\left\{\hat{\xi}'_l(k), \xi_{\min}\right\}, \quad (1)$$

where the parameters w and $\xi_{\min}$ are set to 0.98 and respectively -15 dB. The *a posteriori* SNR is computed as

$$\hat{\gamma}_l(k) = \frac{|Y_l(k)|^2}{\hat{\lambda}_{N_l(k)}}. \quad (2)$$

Note that the clean speech spectral amplitude estimate $|\hat{X}_{l-1}(k)|$ from Eq. 1 is replaced by the actual clean speech spectral amplitude $|X_{l-1}(k)|$, which is known during training.

Next, for each frame l and each frequency bin k both SNR values are uniformly quantized within the range $[-15, 20]$ dB with a stepsize of $\Delta = 1$ dB

$$Q\{\hat{\gamma}_l(k)\} = \tilde{\gamma}_l(k) \rightarrow (i)_{k,l} \in \{1, ..., N_\gamma\},$$
$$Q\{\hat{\xi}_l(k)\} = \tilde{\xi}_l(k) \rightarrow (j)_{k,l} \in \{1, ..., N_\xi\}, \quad (3)$$

where $N_\gamma, N_\xi = 35$. Along with SNR quantizer indices $(i,j)_{k,l}$ the respective *ideal gain* is computed as follows

$$G_l^{id}(k) = \frac{|X_l(k)|^2}{|X_l(k)|^2 + |N_l(k)|^2}. \quad (4)$$

Stepping through all frames l of the training database, a set of ideal gains $\Gamma^{id}_{(i,j),k} = \left\{G_l^{id}(k)\right\}_{\forall l}$ is collected for each frequency bin k and SNR quantizer indices (i,j). The IGA weighting rule $G_{(i,j)}(k)$ for the frequency bin k and SNR quantizer indices (i,j) is obtained by averaging all the gains computed during training [1]

$$G_{(i,j)}(k) = \overline{\Gamma^{id}_{(i,j),k}}\,. \quad (5)$$

The weighting rule on the Bark scale (B-IGA) $G^{Bark}_{(i,j)}(m)$ is calculated for each subband m by averaging the IGA weighting rules for all frequency bins within the corresponding subband

$$G^{Bark}_{(i,j)}(m) = \overline{\left\{G_{(i,j)}(k)\right\}}_{k\in\kappa_m}, \quad (6)$$

where κ_m represents all frequency bins k in subband m, $m \in \{1, \ldots, 19\}$.

It turns out to be advantageous that separate weighting rules are computed during speech presence and speech absence. For each frequency bin, a voice activity detector (VAD) is computed based on *binwise* speech absence probability [10], that is operating on noisy speech. At the end of the training the weighting rules are smoothed by low-pass filtering (e.g. 5×5 Gaussian bell convolution filter) with the purpose of achieving better generalization properties.

Examples of the B-IGA weighting rules trained on car noise are depicted in Figure 1. It is interesting to note that the weighting rule in speech absence does not completely suppress the signal. A non-zero weighting rule value in speech absence can help preserve the speech and noise naturalness, particularly in the transition from speech presence to speech absence, or vice versa. Also of interest is that during speech pause the weighting rule at 1 Bark exhibits lower values than at 14 Bark indicating that the noise is more strongly suppressed at lower frequencies than at higher frequencies. The explanation for this is that the car noise is concentrated mostly in low frequencies.

Another interesting observation is the distribution of SNR indices (i,j) during training. From Figure 1 it can be seen that values $\hat{\gamma}_l(k) < 0$ dB do not occur before LP filtering, since the noise estimator puts an upper bound on the noise spectral variance estimate $\hat{\lambda}_{N_l(k)} \leq |Y_l(k)|^2$. The *a priori* SNR on the other hand is lower bounded by $\xi_{min} = -15$ dB.

2.2. Applying the Weighting Rules

During testing we need to estimate the *a priori* and *a posteriori* SNR values, which are used to retrieve the corresponding gain from the weighting rule table.

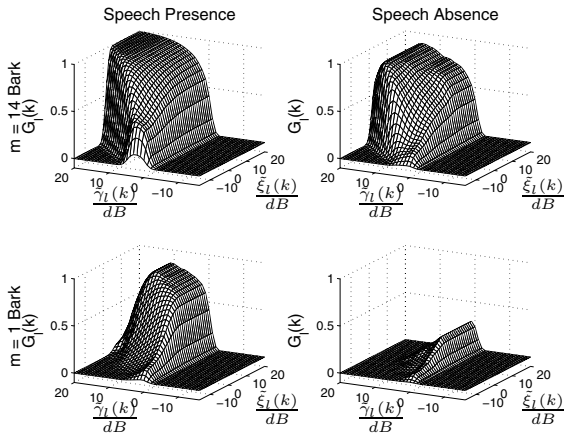

Figure 1: The subband B-IGA weighting rules in speech presence and speech absence: $m = 1$ Bark and $m = 14$ Bark. The weighting rules are trained with car noise.

The *a priori* SNR is computed using $|\hat{X}_{l-1}(k)|$ in Eq. 1 instead of $|X_{l-1}(k)|$. The *a posteriori* SNR is computed exactly like in training. Since the weighting rules for speech presence and absence are different, VAD decision is computed for each frequency bin based on the *a priori* SNR values [10].

Following the SNR quantization described in Eq. 3, both the index pair $(i,j)_{k,l}$ and the VAD decision are used to obtain the appropriate gain value $G_l(k) = G^{Bark}_{(i,j)}(m)$ and estimate the clean speech spectral amplitude $\hat{X}_l(k)$ according to

$$\hat{X}_l(k) = G_l(k) \cdot Y_l(k). \quad (7)$$

2.3. Storage Requirements

Let us compute the memory requirement for IGA and B-IGA. Assuming that each gain value requires 2 Bytes, the IGA approach generally requires $2 \times (\frac{1}{2}K + 1) \times N_\gamma^{\text{eff}} \times N_\xi = 400$ kBytes ($K = 256$, $N_\xi = 35$, $N_\gamma^{\text{eff}} = 22$). Please note that this value is computed by considering only the *effective a posteriori* SNRs N_γ^{eff}, for which the weighting rules assume non-zero values. In contrast to IGA, the newly developed B-IGA method requires only $2 \times M \times N_\gamma^{\text{eff}} \times N_\xi = 58.5$ kBytes where $M = 19$ for the sampling frequency $f_s = 8$ kHz.

3. R-PLS Parameterization

A careful examination of the weighting rules reveals that they have similar shapes. We see from Figure 1 that along the axis $\tilde{\xi}_l(k)$, the weighting rules are monotonically increasing up to a certain point and then "saturate" to a constant.

This pattern can be seen more clearly if we unfold the 2-D weighting rule columnwisely and observe the *weighting gain* [1] for a certain *a posteriori* SNR, as shown in Figure 2.

We notice from Figure 2 that the weighting gain increases monotonically up to a certain point and then saturates. This motivates us to employ an R-order Polynomial Least Square (R-PLS)

[1] the term *weighting gain* will be used to denote the 1-D weighting rule in each *a posteriori* SNR $\tilde{\gamma}_l(k)$

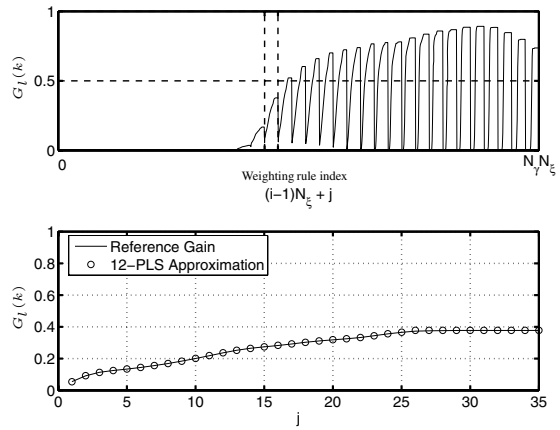

Figure 2: Columnwise weighting rule in speech absence for $m = 14$ Bark (upper) and the corresponding weighting gain for $\tilde{\gamma}_l(k) = 1$ dB (lower).

model for parameterizing the weighting gain. The R-order polynomial function $F_i(x) = \sum_{r=0}^{R} c_{i,r} x^r$ approximates the monotonically increasing curve and the saturation level H^{sat} models the constant part. The R-PLS approximation is performed for each B-IGA weighting rule at the end of the training session in order to compute the parameters $C_i = \{c_{i,0}, \ldots, c_{i,R}\}$ and H_i^{sat}, $i = 1, \ldots, N_\gamma$. Two extra parameters $[j_{a_i}, j_{b_i}]$ are required to indicate the range of the polynomial approximation.

For denoising, the R-PLS parameters are used to reconstruct an approximation of the B-IGA weighting rules $\hat{G}_{(i,j)}^{Bark}(m)$, and the SNR quantizer indices (i, j) are applied for retrieving the appropriate gain value $G_l(k)$. Tables 1 & 2 summarize the parameterization and computation of the R-PLS weighting rules.

The R-PLS parameters take up only $\frac{(R+4)}{N_\xi}$ of the memory required by the B-IGA weighting rules tables. Choosing R between 4 and 12, we can store the weighting rules in 13.5 to 27 kBytes, which represents only 3.3% to 6.7% of the memory required by the IGA weighting rules. In the next section we show that the new weighting rules exhibit virtually no performance loss compared to IGA, although they require 15 to 30 times less storage.

4. Experimental Results

We evaluate the performance of the proposed approach in car noise. In our experiment, 40 different utterances spoken by 8 different speakers (4 male and 4 female) and 84 car noise signals are taken from the NTT-AT speech and noise databases [11, 12]. These signals were split into 2 sets of equal size for training and testing. After combination, $20 \times 42 = 840$ noisy speech utterances at $f_s = 8$ kHz were obtained for each training and testing session.

As a reference system, we employed the *a priori* SNR driven Wiener filter [2] and MMSE Short-Time Spectral Amplitude (MMSE-STSA) [3] with the speech absence probability computed according to [10]. Noise estimation for all compared approaches is being done using the minimum statistics by Martin [9]. The DFT length is $K = 256$, segment/frame length and frame shift is $N = 160$ samples, and window length is 200 samples for both systems.

Figure 3 shows the trained B-IGA weighting rules for speech

For each (*effective*) *a posteriori* SNR index i do:
- Find the *a priori* SNR range $[j_{a_i}, j_{b_i}]$, where the weighting rule $\{G_{(i,j)}^{Bark}(m)\}_{j=j_{a_i}}^{j_{b_i}}$ is monotonically increasing
- Apply Least Squares fitting to compute polynomial coefficients $C_i = [c_{i,0}, \ldots, c_{i,R}]$ that can approximate the weighting rule $\{G_{(i,j)}^{Bark}(m)\}_{j=j_{a_i}}^{j_{b_i}}$
- Compute the saturation level $H_i^{sat} = \max\{G_{(i,j)}^{Bark}(m)\}_{j=1}^{N_\xi}$

Table 1: Summary of the R-PLS parameterization.

For each SNR indices $(i, j)_{k,l}$ do:
- Use index $i_{k,l}$ to address the corresponding R-PLS parameters $\{C_i, H_i^{sat}, j_{a_i}, j_{b_i}\}$
- Compute the polynomial function $F_i(j)_{j=1}^{N_\xi}$ based on the R-PLS coefficients C_i
- Compute the R-PLS approximated weighting rule $\hat{G}_{(i,j)}^{Bark}(m)$

$$\hat{G}_{(i,j)}^{Bark}(m) = \begin{cases} 0 & \text{if } j < j_{a_i} \\ F_i(j) & \text{if } j_{a_i} \le j \le j_{b_i} \\ H_i^{sat} & \text{if } j > j_{b_i} \end{cases}$$

- Use index $j_{k,l}$ to address the R-PLS-approximated weighting rule $G_l(k) = \hat{G}_{(i,j)}^{Bark}(m)$

Table 2: Summary of R-PLS B-IGA weighting rule computation.

presence and absence at 1 Bark, as well as the corresponding 12-PLS parameterization and the approximation error. We see that the 12-PLS parametric representation makes a quasi perfect reconstruction of the B-IGA rule.

The relative performance in terms of segmental noise attenuation and segmental speech-to-speech distortion ratio [1] is illustrated in Figure 4 for the Wiener filter, MMSE-STSA, original IGA, as well as B-IGA and the associated R-PLS approximation for $R = \{4, 8, 10, 12\}$. The more a curve is located in the upper right, the less residual noise and speech distortion remain, and the better the algorithm performs.

From the results we draw the conclusion that the environment-dependent speech enhancement performs better than both the Wiener filter and MMSE-STSA, which is also confirmed by informal listening tests. We see that B-IGA gives the same performance as IGA, and that the 12-PLS parameterization of B-IGA is indistinguishable from IGA with only 6.7% of the memory required by the original algorithm IGA weighting rules.

Using values for R lower than 12 in the R-PLS approximation we get a trade-off between performance and storage requirements. For example 4-PLS gives results comparable to IGA for 30 times less storage (i.e. 3.3%) and is still better than Wiener filtering and MMSE-STSA.

5. Conclusion

Environment-dependent weighting rules for speech enhancement trained on a specific noise type outperform well-known state-of-the-art environment-independent techniques, such as Wiener filtering or MMSE-STSA. One deficiency of the original approach, termed IGA, is the relatively high cost in terms of memory for

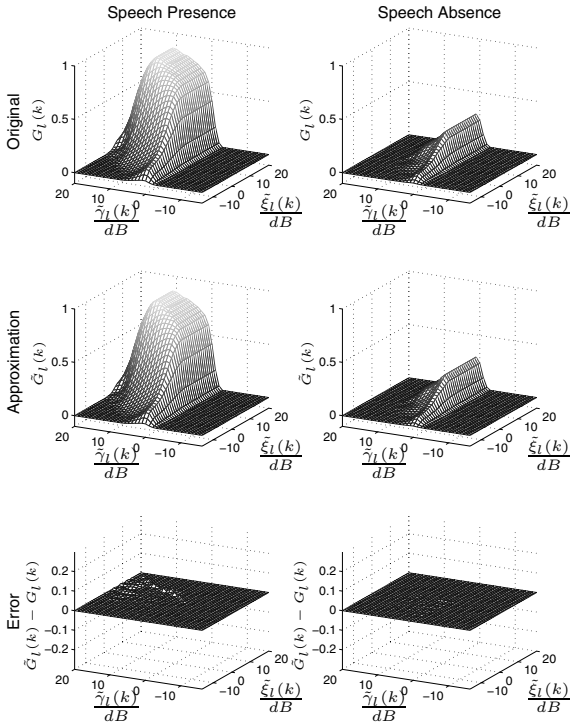

Figure 3: 12-PLS approximation of the B-IGA weighting rule in speech presence and in speech absence for $m = 1$ Bark.

storing the trained weighting rules, which makes it unattractive for embedded DSP applications.

In this paper we presented a new environment-dependent speech enhancement method, which requires 15 to 30 times less storage than IGA with virtually no loss in performance. The new approach redefines the weighting rules on the Bark scale and stores a parametric representation of them obtained by polynomial modelling of the unfolded gains.

6. References

[1] T. Fingscheidt and S. Suhadi, "Data-Driven Speech Enhancement," in *Proc. of ITG-Fachtagung "Sprachkommunikation"*, Kiel, Germany, Apr. 2006, VDE–Verlag.

[2] P. Scalart and J.V. Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP'96*, Atlanta, GA, May 1996, pp. 629–632.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

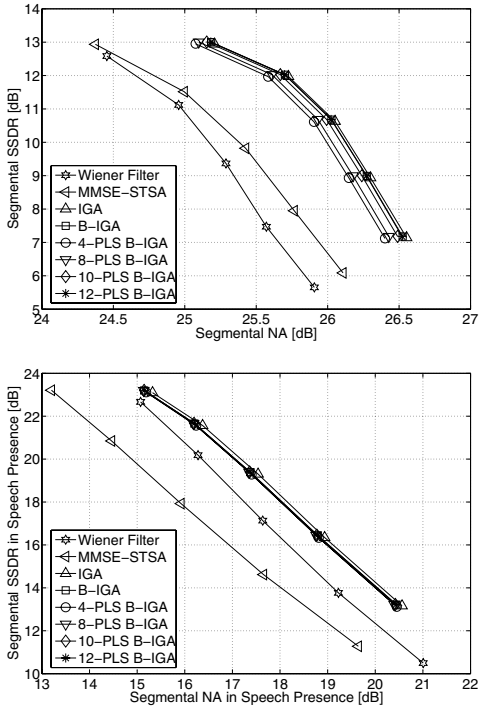

Figure 4: Segmental SSDR vs. noise attenuation in the whole utterance (above) and in speech presence (below) .

[5] P.C. Loizou, "Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, Sept. 2005.

[6] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors," in *Proc. of ICASSP'02*, Orlando, Florida, May 2002, pp. 504–512.

[7] T. Lotter and P. Vary, "Noise Reduction by Maximum a Posteriori Spectral Amplitude Estimation with Supergaussian Speech Modeling," in *Proc. of IWAENC*, Kyoto, Japan, Sept. 2003, pp. 83–86.

[8] J.E. Porter and S.F. Boll, "Optimal Estimators for Spectral Restoration of Noisy Speech," in *Proc. of ICASSP'84*, San Diego, California, Mar. 1984, pp. 18A.2.1–18A.2.4.

[9] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[10] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 112–116, Apr. 2002.

[11] NTT-AT Speech Database, "Multi-Lingual Speech Database for Telephonometry 1994," `http://www.ntt-at.com/products_e/speech/index.html`, 1994.

[12] NTT-AT Noise Database, "Ambient Noise Database for Telephonometry 1996," `http://www.ntt-at.com/products_e/noise-DB/index.html`, 1996.