On the Enhancement of Dereverberation Algorithms Based on a Perceptual Evaluation Criterion

Thiago de M. Prego1,2, Amaro A. de Lima1,2, and Sergio L. Netto2

Program of Electrical Engineering, Federal University of Rio de Janeiro, Brazil. Program of Electrical Engineering, Federal Center for Technological Education, Brazil.

{thiago.prego, amaro.lima, sergioln}@smt.ufrj.br

Abstract

This paper analyzes and proposes an optimization procedure for dereverberation algorithms based on a perceptual assessment criterion for dereverberation algorithms. The complete procedure is applied to the adaptive inverse-filtering (AIF) and spectral subtraction (SS) stages of a given dereverberation algorithm using the so-called QAverb perceptual quality measure. Experimental results, using a 204-signal speech database, indicate that the associated algorithm can be greatly simplified (in about 97% of the overall computational complexity) by removing the AIF stage. In addition, a fine-tuning of the SS stage is able to improve in 6% the algorithm’s QAverb score, resulting in a much simpler and more efficient algorithm in a perceptual point of view.

1. Introduction

Reverberation can strongly affect the performance of state-of-the-art systems of speech/speaker recognition and hearing-aid, motivating the use of speech enhancement techniques. Although reverberation can be damaging to speech intelligibility and perceptual quality, in a small amount it makes speech more pleasant to common listeners [1]. The use of a microphone array is commonly associated to dereverberation techniques, but for the applications previously mentioned the use of one microphone approach is more indicated.

This paper analyzes and proposes an optimization procedure for dereverberation algorithms based on a perceptual assessment objective measure [2]. The entire procedure is applied to a two-stage one-microphone algorithm for reverberant speech enhancement introduced in [3] consisting of two isolated signal-processing blocks (hereby referred to as the AIF and SS stages), as illustrated in Fig. 1, where y(n), z(n), and x(n) are the reverberant, inverse-filtered, and spectral-subtracted (or dereverberated) signals, respectively.

The main concept behind this algorithm comes from the commonly adopted model for the reverberant room impulse response (RIR), which is composed by three parts: the direct path signal, which corresponds to the direct path speech from the source to the listener; the early reflections, which presents a non-flat frequency response that distorts the speech spectrum; and finally the late reverberation, which causes smearing of the speech spectrum, reducing the intelligibility and quality of the signal [3]. This two-stage algorithm was conceived to mitigate the effects due to the early and late reflections, which are associated with coloration and long-term reverberation [4], respectively, as detailed in the following subsections.

2. A Two-Stage Dereverberation Algorithm

The algorithm introduced [3] consists of two isolated signal-processing blocks (hereby referred to as the AIF and SS stages), as illustrated in Fig. 1, where y(n), z(n), and x(n) are the reverberant, inverse-filtered, and spectral-subtracted (or dereverberated) signals, respectively.

Figure 1: Diagram of the two-stage algorithm.

The main concept behind this algorithm comes from the commonly adopted model for the reverberant room impulse response (RIR), which is composed by three parts: the direct path signal, which corresponds to the direct path speech from the source to the listener; the early reflections, which presents a non-flat frequency response that distorts the speech spectrum; and finally the late reverberation, which causes smearing of the speech spectrum, reducing the intelligibility and quality of the signal [3]. This two-stage algorithm was conceived to mitigate the effects due to the early and late reflections, which are associated with coloration and long-term reverberation [4], respectively, as detailed in the following subsections.

2.1. Inverse filtering

The main objective of the AIF block is to reduce the coloration effect by reconstructing an estimate of the original (clean) speech signal. This block is based on [5], where a multi-

Copyright © 2013 ISCA
microphone setup is used to determine an RIR estimate by maximizing the kurtosis of the linear prediction (LP) residue. Therefore, an inverse filter for the estimated RIR is applied to the reverberant speech to obtain a cleaner signal.

In this context, the inverse-filtered speech \( z(n) \) can be described as

\[
z(n) = h_y(n) + y(n),
\]

where \( h_y(n) = \sum_{j=0}^{L-1} g_y(n-j) \) is the RIR.

The \( K \)-length inverse filter is then given by \( g = [g_0, g_1, \ldots, g_{K-1}]^T \), where \( g \) is designed to maximize the kurtosis of \( z(n) \) using, for instance, a length-\( K \) block least-mean-squares (LMS) adaptive algorithm. Instead of using \( g(n) \), however, the adaptive algorithm uses the block version of the LP residue defined by \( y(m(\frac{K}{2})) := y(m(\frac{K}{2}) + K - 1) \). Equivalently the \( m \)-th block of inverse-filtered LP residue \( z_r(m) \) is generated from the block version of Eq. (1).

The adaptive algorithm uses the average kurtosis of \( z_r(m) \), as cost function:

\[
J = \frac{1}{M} \sum_{m=0}^{M-1} J(m) = \frac{1}{M} \sum_{m=0}^{M-1} \left( \frac{\mathbb{E}[z_r^4(m)]}{\mathbb{E}[z_r^2(m)]^2} - 3 \right),
\]

where \( \mathbb{E}[\cdot] \) denotes the statistical mean operator, such that \( f(m) = \mathbb{E}[f(m)] \).

To avoid a large eigenvalue spread of the input-signal autocorrelation matrix, which would lead to slow or no convergence, a frequency-domain adaptive algorithm may be employed by applying a fast Fourier transform (FFT) to all length-\( K \) data blocks [6]. Thus, by defining \( G(i) \) as the FFT of \( g(i) \) from the \( i \)-th iteration, \( F(m) \) and \( Y_r(m) \) being respectively the FFTs of \( f(m) \) and \( y_r(m) \), the AIF update equation becomes

\[
G(i + 1) = G(i) + \mu \sum_{m=0}^{M-1} F(m) Y^*_r(m),
\]

where \( \mu \) is the adaptive-filter step size and the superscript asterisk represents the complex-conjugate operation. Once the optimum filter \( g \) is obtained, the inverse-filtered speech \( z(n) \) is calculated according to Eq. (1).

Far all practical issues, the algorithm described in [3] uses an adaptation step size \( \mu = 3 \times 10^{-3} \), an LP filter length \( K = 10 \), and a block size \( K = 0.032 \times F_s \), with 50\% overlap between consecutive blocks, \( F_s \) being the sampling frequency. Once the optimum filter \( g \) is obtained, the inverse-filtered speech \( z(n) \) is calculated according to Eq. (1).

Far all practical issues, the algorithm described in [3] uses an adaptation step size \( \mu = 3 \times 10^{-3} \), an LP filter length \( K = 10 \), and a block size \( K = 0.032 \times F_s \), with 50\% overlap between consecutive blocks, \( F_s \) being the sampling frequency. Once the optimum filter \( g \) is obtained, the inverse-filtered speech \( z(n) \) is calculated according to Eq. (1).

The SS block, as detailed in Fig. 2, aims at the reduction of the late reverberation component of the RIR. The SS stage starts with the inverse-filtered speech \( z(n) \) and outputs the dereverberated speech \( x(n) \), whose phase is derived directly from \( z(n) \).

Let \( S_r(k,m) = |S_r(k,m)|e^{j \varphi_r(k,m)} \) be the FFT of the \( m \)-th frame of the windowed version of \( z(n) \), using a 32 ms hamming window with 24 ms overlap between consecutive frames. Let also \( \rho \) be the length of the early reflection in frames, commonly considered around 50 ms, corresponding to \( \rho = 7 \) with \( F_s = 48 \) kHz; \( \gamma \) be the scaling factor that establishes the relative strength of the late impulse components after the inverse filtering being set to \( \gamma = 0.35 \); and \( w(m) \) be an asymmetrical smoothing window based on the Rayleigh distribution, given by

\[
w(m) = \begin{cases} \frac{m-a}{a} e^{-\frac{(m-a)^2}{2a^2}}, & \text{if } m > -a \\ 0, & \text{otherwise} \end{cases},
\]

where \( a < \rho \) controls the overall function time spread. In the original configuration, \( a = 5 \) frames, thus providing a reasonable match to the shape of the equalized impulse response.

The model of the power spectrum of the late reverberation, in order to perform the spectral subtraction, can be described as

\[
|S_r(k,m)|^2 = |S_s(k,m)|^2 + \epsilon |S_z(k,m)|^2,
\]

where \( k, m \) are the frequency- and frame-domain indexes, respectively.

Assuming that the early and late components are mutually uncorrelated, the power spectrum of the early impulse components can be estimated by subtracting the power spectrum of the late impulse components from the inverse-filtered speech. The spectrum subtraction scheme performs a weighting in the power spectrum of \( z(n) \), where the block SUBTRACTION is given by

\[
|S_e(k,m)|^2 = \max \left[ 1 - \frac{|S_s(k,m)|^2}{|S_z(k,m)|^2}, \epsilon \right],
\]

where \( \epsilon = 0.001 \) corresponds to the maximum attenuation of 30 dB, and finally the magnitude spectrum of \( z(n) \) is given by

\[
|S_e(k,m)| = \sqrt{|S_s(k,m)|^2 \times |S_z(k,m)|^2}.
\]

In order to calculate \( z(n) \), the phase \( \varphi_s(k,m) \) of \( S_e(k,m) \) is combined to the magnitude \( |S_e(k,m)| \), such that

\[
S_e(k,m) = |S_e(k,m)| e^{j \varphi_s(k,m)}.
\]

3. Proposed Modifications

### 3.1. Perceptual Quality measurements

In order to assess the perceptual quality of a reverberant speech signal, this work employs the QAreverb [2] measure \( Q \), defined as

\[
Q = -\frac{T_{60} \sigma^2}{R \gamma},
\]

where \( T_{60} \) is the reverberation time (defined as the period of time required for the sound-pressure to decay 60 dB, which in this work is estimated by Karjalainen’s algorithm [7]), \( \sigma^2 \) is the room spectral variance (RSV) [8], and \( R \) is the direct-to-reverberant energy ratio (DRR) [9], and \( \gamma = 0.3 \). In practice, a higher \( T_{60} \) indicates a more lasting reverberation effect, the
RSV is closely related to the coloration effect, and the DRR provides some insight on the source-microphone relative position. These three measures can be obtained directly from the RIR, $h(n)$, which is estimated from the deconvolution process between the clean and reverberant speech signals.

In practice, the QAreverb measure can be seen as an extension of Allen’s score [10] by incorporating the DRR into its formulation. The $Q$ score is mapped onto the $1−5$ mean opinion score (MOS) scale through a third order polynomial (followed by a first-order polynomial mapping for some bias adjustment), as detailed in [2].

### 3.2. Reverberant speech database

The main database used in this work is called the new Brazilian-Portuguese (NBP) database [2] which uses a $F_s = 48$-kHz sampling frequency. In this database, 4 anechoic speech signals (2 by male speaker and 2 by female speaker) were used to generate reverberant speech following three different frameworks:

(i) Artificial reverberation: This scenario is represented by 6 distinct artificially generated RIRs, where the early reflections were modeled via the image method, and the late reverberation used the feedback delay network method and a modified version of Gardner’s method, for emulating the lower and higher reverberation times, respectively. The average $T_{60}$ for the 6 RIRs are given by [196, 292, 387, 469, 574, 664] ms.

(ii) Natural reverberation: This approach consists in 17 different RIRs obtained from the direct recordings of 4 different types of rooms with several source-microphone distances for each room, as detailed in [14]. The average $T_{60}$ for each of the 4 rooms is in the range of [120, 230, 450, 780] ms.

(iii) Real reverberation: In this case, the degraded signals were directly played/recorded in 7 real rooms and 4 different types of rooms, as detailed in [13]. The average $T_{60}$ for each of the 11 rooms are in the range of [140, 390, 570, 650, 790, 890, 920] ms.

As described in [2], the perceived quality of all 204 NBP signals (4 anechoic, 24 artificial, 68 natural, and 108 real cases) was assessed through an absolute category rate (ACR) MOS test with 30 non-trained listeners for each signal. 10 additional signals covering the whole NBP reverberation range were used in the initial part of the test to assist the listener in adjusting his/her scoring scale. The scores of these 10 additional signals were discarded later on without the listener’s knowledge. In the end, outliers (scores outside the region of three standard deviations around the mean score of each signal) were removed. Only 9, all from different listeners and signals, out of a total of 6120 scores were removed in this procedure. The NBP database is available upon request by e-mail to the authors.

The NBP database was divided into two sub-databases: the training database, composed of 18 reverberant speech signals, one for each environment (1 anechoic, 6 artificial RIRs, 4 natural rooms, and 7 real rooms). The training database is composed of the remaining 182 reverberant speech signals from the NBP database. Initially the parameters were optimized regarding $Q_{MOS}$ in the training database and then the test database is used to validate the system performance.

### 3.3. Methodology

For a complete fine-tuning of the dereverberation algorithm, three experiments were devised, as detailed below:

**Experiment 1:** In a first step, the optimization of the AIF block parameters LP-filter length $K$ and adaptation step $\mu$, as detailed in Section 2.1, is performed with the SS block set to its original configuration. Furthermore, a new convergence criterion is considered to accelerate the entire adaptation process [12].

**Experiment 2:** In a second experiment, following the SS description provided in Subsection 2.2, the optimized parameters are the scaling factor $\gamma$, the attenuation limit $\epsilon$, the length of the early reflections $\rho$, and the spread control $a$. In this scenario, the AIF block is set as in the original algorithm design.

**Experiment 3:** In a final stage, the two AIF parameters $(K$ and $\mu$), along the new convergence criterion, are jointly optimized with the four SS parameters $(\gamma, \epsilon, \rho, a)$ considered in Experiment 2 above. The best performance settings are then applied to the test database for validation of the overall experimental procedure, as detailed below.

### 4. Experimental Results

#### 4.1. Experiment 1

Since higher LP orders $K$ tend to provide less estimation residues, generating a more impulsive-like profile related to the glottal pulses, this experiment considers the influence of $K$ in the perceived quality of the dereverberated speech signals, along with the influence of the adaptation step size $\mu$.

With the objective of improving the number of iterations for the update of $G$, a new stop criterion was devised for the adaptation algorithm. For this matter, consider the average kurtosis variation in time given by

$$J_d(i) = \frac{\sum_{l=1}^{M} \bar{J}(i-l) - \sum_{l=1}^{M} \bar{J}(i-l+1)}{\sum_{l=1}^{M} \bar{J}(i)}.$$  \tag{10}

By using a stop criterion of the form $J_d(i) = J_{d,\text{max}}$, with for instance $M = 4$, one is able to decrease the average number of iterations significantly, without affecting the perceived quality of the dereverberated speech signals [12].

The AIF optimization considered the parameter ranges $K = \{10, 20, \ldots, 100\}$, $\mu = \{1, 2, \ldots, 10\} \times 10^{-9}$, and $J_{d,\text{max}} = \{-\infty, -100, -75, -50, -25\}$, which gives a total of 500 training setups. In this experiment the parameters of the spectral subtraction block were kept as in [3]. The parameters were analyzed considering the average $Q_m$ for all $Q_{MOS}$ scores within the 18-signal training dataset. The best mean score $Q_m = 3.71$ was obtained by using $K = 40, \mu = 4 \times 10^{-9}$, and $J_{d,\text{max}} = -25$ dB, and the average number of iterations was reduced from 500 to 50.

The 19 best $(K, \mu, J_{d,\text{max}})$ sets with respect to the average $Q_m$, along with the original set, were then selected for the joint optimization stage implemented in Experiment 3.

#### 4.2. Experiment 2

In this scenario, the optimization of the four SS parameters was performed in two pair, while keeping the AIF parameters as set in [3].

First, a joint search for the best values of $\gamma$ and $\epsilon$ was done, and the 14 $(\gamma, \epsilon)$ pairs with the best $Q_m$ scores, along with the original $(\gamma = 0.35, \epsilon = 10^{-3})$ pair, were selected for the joint $(\gamma, \epsilon, \rho, a)$ optimization.

Later, a joint search for the best values of $\rho$ and $a$ was performed, and the 14 $(\rho, a)$ pairs with the best $Q_m$ scores, along with the original $(\rho = 7, a = 5)$ pair, were selected for the joint $(\gamma, \epsilon, \rho, a)$ optimization.
Finally, all 15 \{\gamma, \epsilon\} pairs were combined with the 15 \{\rho, a\} pairs previously chosen, totaling 225 \{\gamma, \epsilon, \rho, a\} distinct setups. The best perceptual score achieved in this experiment was \(Q_m = 3.61\) for the set \(\{\gamma = 0.35, \epsilon = 10^{-3}, \rho = 7, a = 6\}\).

The 19 best sets \{\gamma, \epsilon, \rho, a\} with the best \(Q_m\) scores, along with the original \(\{\gamma = 0.35, \epsilon = 10^{-3}, \rho = 7, a = 5\}\) set, were then selected for the joint optimization procedure implemented in Experiment 3.

### 4.3. Experiment 3

This scenario considers the joint AIF and SS optimization by combining the best 20 parameter sets identified in each of the Experiments 1 and 2 above, leading to a total of 400 distinct setups. After testing all these 400 configurations, the best perceptual performance \(Q_m = 3.71\) was achieved by the set \(\{K = 40, \mu = 4 \times 10^{-9}, J_m^{\max} = -25\,\text{dB}, \gamma = 0.35, \epsilon = 10^{-3}, \rho = 7, a = 6\}\), which corresponds to the same \(Q_m\) performance obtained at the AIF optimization stage. In order to understand why the joint optimization improve upon the results in Experiment 1, the QAreverb metric was broke down into its three partial metrics, as detailed in Tables 1 and 2. From these tables, one clearly notices how the joint optimization does not improve significantly upon the joint optimization in any aspect whatsoever.

#### Table 1: Average performance of the individual reverberation metrics for the training database in Experiment 3.

<table>
<thead>
<tr>
<th>Quality metric</th>
<th>Unprocessed database</th>
<th>Two-stage algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Q_m)</td>
<td>3.46</td>
<td>3.46</td>
</tr>
<tr>
<td>(T_{10})</td>
<td>440</td>
<td>354</td>
</tr>
<tr>
<td>(\sigma_r^2)</td>
<td>5.50</td>
<td>6.95</td>
</tr>
<tr>
<td>(R)</td>
<td>5.54</td>
<td>1.97</td>
</tr>
</tbody>
</table>

This conclusion can be associated to the fact that the AIF stage is somehow degrading the signal in a manner that the SS cannot compensate for. To confirm this assumption, the AIF block was removed from the overall algorithm and the SS stage was once again optimized, following the same strategy as above (Experiment 2). The results for this new optimization round are seen in Tables 3 and 4, for the training and testing datasets, respectively. These new results indicate that the new algorithm, which includes only the SS stage, is able to achieve better perceptual scores (either through better individual metrics or with a combined QAreverb score) compared to any previous algorithm configuration, which was corroborated by informal subjective tests. An additional advantage of the simplified algorithm is that its processing time was further reduced in about 30%, by completely eliminating the AIF stage, in comparison to the already improved configuration provided in Experiment 1.

#### Table 3: Average performance of the individual reverberation metrics for the training database using modified algorithm.

<table>
<thead>
<tr>
<th>Quality metric</th>
<th>Unprocessed database</th>
<th>SS Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Q_m)</td>
<td>3.36</td>
<td>3.65</td>
</tr>
<tr>
<td>(T_{10})</td>
<td>440</td>
<td>455</td>
</tr>
<tr>
<td>(\sigma_r^2)</td>
<td>5.50</td>
<td>5.41</td>
</tr>
<tr>
<td>(R)</td>
<td>5.54</td>
<td>5.61</td>
</tr>
</tbody>
</table>

#### Table 4: Average performance of the individual reverberation metrics for the test database using modified algorithm.

<table>
<thead>
<tr>
<th>Quality metric</th>
<th>Unprocessed database</th>
<th>SS Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Q_m)</td>
<td>3.36</td>
<td>3.53</td>
</tr>
<tr>
<td>(T_{10})</td>
<td>517</td>
<td>490</td>
</tr>
<tr>
<td>(\sigma_r^2)</td>
<td>5.61</td>
<td>5.70</td>
</tr>
<tr>
<td>(R)</td>
<td>7.59</td>
<td>6.12</td>
</tr>
</tbody>
</table>

### 6. Acknowledgements

The authors would like to thank Prof. M. Karjalainen, for providing the \(T_{10}\) estimation algorithm, and Prof. D. Wang, for providing the original dereverberation algorithm.

### 7. References


