Perceptual Improvement of a Two-Stage Algorithm for Speech Dereverberation

Thiago de M. Prego[1], *Amaro A. de Lima*[1,2] *and Sergio L. Netto*[1]

[1]Program of Electrical Engineering, COPPE, Federal University of Rio de Janeiro, Brazil.
[2]Federal Center for Technological Education Celso Suckow da Fonseca (CEFET-RJ), Brazil.
{thprego, amaro, sergioln}@lps.ufrj.br

Abstract

This paper presents three effective proposals for a two-stage algorithm for one-microphone reverberant speech enhancement. The original algorithm is divided into two blocks: one that deals with the coloration effect, due to the early reflections, and the other for reducing the long-term reverberation. The proposed modifications consider changing the linear-prediction model order, the adaptation stepsize and stop criterion for the first algorithm stage. All the modifications are evaluated by a perceptual-quality measure specific for the speech-reverberation context. Experimental results for a 200-signal database show that the proposed improvements yielded an increase of 12% in perceptual measure and a reduction of about 96% in the computation cost when compared to the original framework.

1. Introduction

Speech intelligibility and quality are affected by several kinds of impairments during signal generation, processing or transmission. Such impairments, may include, for example, speech coding distortions, packet loss, time clipping, background noise, echo and reverberation. Although one may prefer most of these impairments to be absent, the reverberation in a small amount turns the speech more pleasant [1] for normal listeners. However reverberation can drastically affect the performance of current automatic speech/speaker recognition or hearing-aid systems, thus requiring an appropriate speech enhancement technique to reduce its perceived effects. Common dereverberation techniques use microphone arrays, but for the applications previously mentioned the use of one-microphone system seems to be more natural and adequate.

In this paper, analysis and improvement proposals are suggested to a two-stage one-microphone algorithm for reverberant speech enhancement [2]. This algorithm is divided into two parts: the first deals with problem of coloration of reverberant speech due to early reflection and the other part treats the long-term reverberation effect. Coloration is often mitigated by an adaptive inverse-filter procedure which maximizes some statistics of the linear prediction (LP) residue to reconstruct the desired speech signal. Meanwhile, the diffuse nature of the late reverberation is often dealt with a spectral subtraction procedure

To introduce the proposed modifications, this paper is organized as follows: In Section 2, the original dereverberation algorithm is explained in details. Section 3 considers three different modifications to the original algorithm in order to increase its dereverberation performance and reduce its computational cost. Section 4 describes a 200-signal database employed in this work and analyzes in details the results achieved by the three suggested improvements for the entire database. Finally, a conclusion concerning the performance increase and the computational reduction is included in Section 5.

2. Original two-stage algorithm

The reverberation effect is often modeled by a room impulse response (RIR) which includes three different portions: the direct-path signal, which corresponds to the direct sound component 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 [2, 3].

The algorithm under analysis was designed to mitigate the effects due to the early and late reflections, which are known as coloration and long-term reverberation, respectively. The two-stage dereverberation algorithm described in [2] consists in applying two isolated signal processing blocks to reduced the reverberation level in a given speech sample. These two blocks are seen in Figure 1, where $y(n)$, $z(n)$ and $x(n)$ are the reverberant, inverse-filtered and spectral-subtracted/dereverberated speech signals, respectively. As all proposed modifications considered in this paper refer to the inverse-filtering block, this stage is described in details in the following subsection.

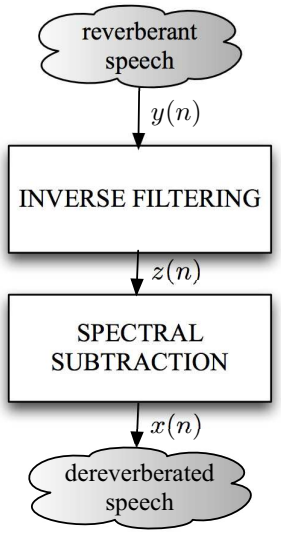

Figure 1: *Block diagram of dereverberation algorithm proposed in [2].*

2.1. Inverse filtering

The idea of the inverse filtering is to reconstruct a estimate of the original speech signal, reducing the effects of coloration. This block was based on [4], where a multimicrophone inverse filtering algorithm is determined by maximizing the kurtosis of the LP residue generating an inverse model for the associated reverberation RIR.

Defining the inverse filter of length L as $\mathbf{g} = [g_0, g_1, \ldots, g_{L-1}]^T$, whose impulse response is $h_g(n) = \sum_{j=0}^{L-1} g_j \delta(n-j)$, the inverse-filtered $z(n)$ speech may be written as

$$z(n) = h_g(n) * y(n). \quad (1)$$

As the LP residue of clean speech has higher kurtosis than of reverberant speech, the inverse filter $\mathbf{g}$ may be designed to maximize the kurtosis of $z(n)$. This optimized design may be implemented in an adaptive manner, based on the problem of LP reconstruction artifacts [4], which uses the LP residue $y_r(n)$ of the reverberant speech instead of $y(n)$ in Eq. (1) leading to $z_r(n) = h_g(n) * y_r(n)$. Once one obtains the optimum filter $\mathbf{g}$, the inverse-filtered speech $z(n)$ is calculated according to Eq. (1).

The optimization of $\mathbf{g}$ is calculated based on a length-$\bar{L}$ block Least Mean Squares (LMS)-like adaptive algorithm, which exploits the mth block of $y_r(n)$, defined by $\mathbf{y_r}(m) = [y_r(m(\frac{3}{2}\bar{L})), \ldots, y_r((m+1)(\frac{3}{2}\bar{L})-1)]$. Equivalently the mth block of inverse-filtered LP residue $\mathbf{z_r}(m)$ is generated. The adaptive algorithm uses the average kurtosis of $\mathbf{z_r}(m)$, as cost function:

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

where $\mathbb{E}[\cdot]$ denotes the statistical mean operator, such that

$$\begin{aligned}\mathbf{f}(m) &= \nabla J(m) \\ &= \frac{4(\mathbb{E}[{\mathbf{z_r}}^2(m)]{\mathbf{z_r}}^3(m) - \mathbb{E}[{\mathbf{z_r}}^4(m)]\mathbf{z_r}(m))}{\mathbb{E}^3[{\mathbf{z_r}}^2(m)]}. \quad (3)\end{aligned}$$

According to Haykin [5] the time-domain implementation is not recommended for this formulation, as a large eigenvalue spread of the input-signal autocorrelation matrix may cause slow or no convergence. Thus a frequency-domain structure is employed, where the FFT (Fast Fourier Transform) is applied to length-$\bar{L}$ data blocks. Defining $\mathbf{G}(i)$ as the FFT of $\mathbf{g}^{(i)}$ from ith iteration, $\mathbf{F}(m)$ and $\mathbf{Y_r}(m)$ being respectively the FFT's of $\mathbf{f}(m)$ and $\mathbf{y_r}(m)$, the inverse filter update equation becomes

$$\mathbf{G}(i+1) = \mathbf{G}(i) + \frac{\mu}{M}\sum_{m=0}^{M-1}\mathbf{F}(m)\mathbf{Y}_\mathbf{r}^*(m), \quad (4)$$

where μ is the adaptive filter stepsize and the superscript asterisk represents the complex-conjugate operation. Considering practical issues, the algorithm described in [2] uses an adaptation stepsize $\mu = 3 \times 10^{-9}$, an LP filter length $L = 10$ and a block size $\bar{L} = 0.032 \times F_s$ with 50% of overlap between consecutive blocks, where F_s is the sampling frequency.

3. Proposed modifications

This work considers three design aspects within the inverse-filtering block:

1. The LP filter order;
2. The influence of the adaptation step; and
3. The convergence criterion for the adaptive algorithm.

The perceptual quality of the resulting dereverberated speech may be evaluated by Allen's objective measure defined as [6, 7]

$$P = P_{max} - \sigma_V^2 T_{60}, \quad (5)$$

where P_{max} is the maximum possible score, σ_V^2 is the room spectral variance [9] and T_{60} is the associated reverberation time. In practice, σ_V^2 is highly associated to the coloration effect and T_{60} indicates a more lasting reverberation effect, and these two values may be obtained directly from the RIR, $h(n)$, which is estimated by a deconvolution process between the clean and the degraded speech signals. The Allen's measure presented in Eq. (5) with $P_{max} = 0$ is used in this work to evaluate the quality of the dereverberated speech.

Initially the proposed approaches were analyzed observing just one reverberant speech signal degraded by real reverberation effect with 960 ms of reverberation time, which is expected to represent a general environment reverberant situation. A subsequent evaluation for a complete 200-signal database is presented in Section 4.

3.1. Influence of LP filter order

The analysis of the LP filter order was first motivated by observing the behavior of the average kurtosis $\bar{J}(i)$, which depends on the iterations i, since $\mathbf{z_r}(m)$ is updated by $\mathbf{g}^{(i)}$, as depicted in Figure 2, which, after convergence, seems to oscillate around a mean value and generating quite distinct dereverberation performances in each case.

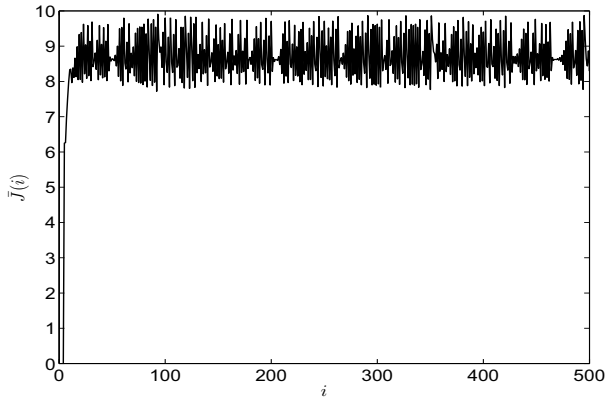

Figure 2: *Average kurtosis $\bar{J}(i)$ convergence using $L = 10$ LP coefficients, $\mu = 3 \times 10^{-9}$ and a total of $N_i = 500$ iterations.*

The increase in LP filter order should provide a more impulsive-like LP residue in accordance to the glottal cycles, which is a more appropriated structure to be optimized by the kurtosis maximization. Figure 3 presents the results of Allen's score (P) of the dereverberated speech varying the LP filter order from 10 to 100 coefficients, using the original adaptation step $\mu = 3 \times 10^{-9}$ and $N_i = 500$ iterations. From this figure, the best score was achieved by $L = 50$ coefficients, corresponding to an approximate 12% increase in P and the smoother kurtosis profile depicted in Figure 4.

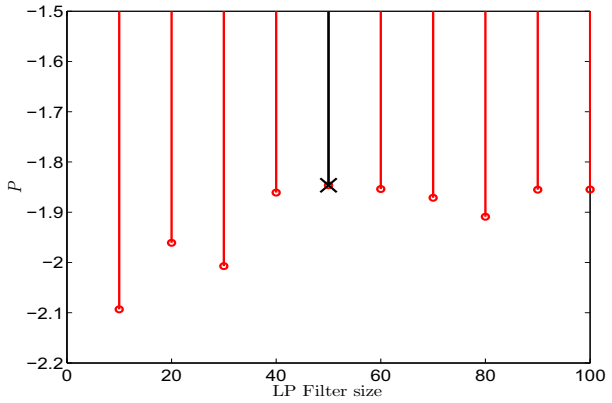

Figure 3: *Dereverberated speech quality applying Allen's score (P) as a function of the LP filter order and with* $\mu = 3 \times 10^{-9}$ *and a total of* $N_i = 500$ *iterations.*

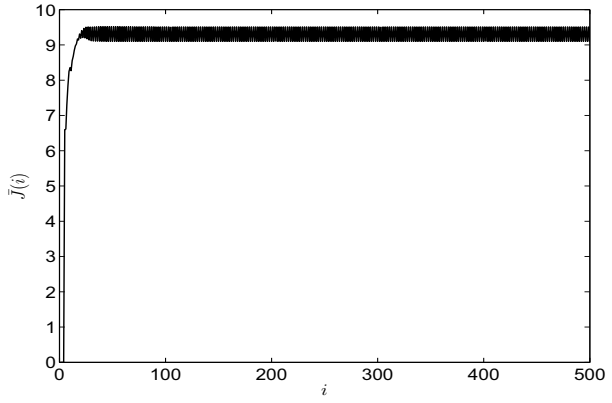

Figure 4: *Average kurtosis* $\bar{J}(i)$ *convergence using* $L = 50$ *LP coefficients,* $\mu = 3 \times 10^{-9}$ *and a total of* $N_i = 500$ *iterations.*

3.2. Influence of adaptation stepsize (μ)

Figure 5 depicts the Allen's score (P) for several values of μ ranging from 1×10^{-9} to 1.4×10^{-8}, with $L = 50$ and $N_i = 500$. Following this analysis, the chosen adaptation step was set to $\mu = 7 \times 10^{-9}$ (best performance), which yields an additional 20% increase in Allen's score, as compared to the results in the previous subsection.

3.3. Convergence criterion for the adaptation algorithm

In order to avoid an excessive computational cost for all $N_i = 500$ iterations, a new stopping criterion was devised for the adaptation algorithm, thus reducing the overall computational burden while sustaining a similar quality of the dereverberated speech. In this case, we measure the average kurtosis variation in time by

$$\bar{J}_d(i) = \frac{\left|\sum_{l=1}^{\bar{M}} \bar{J}(i-l) - \sum_{l=1}^{\bar{M}} \bar{J}(i-l+1)\right|}{\left|\sum_{l=1}^{\bar{M}} \bar{J}(i-l)\right|} \qquad (6)$$

and set the stopping criterion at $\bar{J}_d(i) = -50$ dB, with $\bar{M} = 4$ obtained empirically to yield an smooth process. The resulting kurtosis variation for $L = 50$ and $\mu = 7 \times 10^{-9}$ is seen in Figure 6, where one notices a approximately flat pattern with

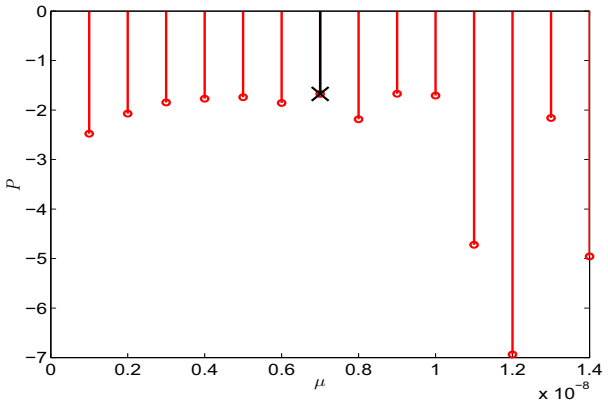

Figure 5: *Dereverberated speech quality using Allen's score (P) as a function of the adaptation stepsize (*μ*) with* $L = 50$ *LP coefficients and a total of* $N_i = 500$ *iterations.*

minimums up to -80 dB and a threshold reach after only $N_i = 29$ iterations, which represents a computational-cost decrease of approximately 94% in the inverse-filter update process.

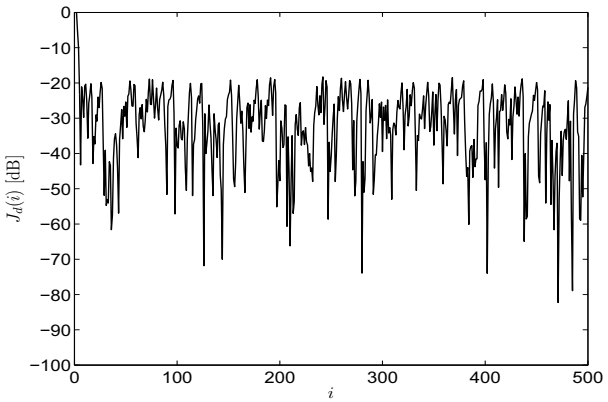

Figure 6: *Average kurtosis variation with* $\bar{M} = 4$*,* $L = 50$ *LP coefficients and* $\mu = 7 \times 10^{-9}$.

4. Analysis for complete database

The database used in this work is called the New Brazilian-Portuguese (NBP) database and includes 200 signals, with $F_s = 48$ kHz, and different types and levels of reverberation. The whole database was generated from 4 anechoic speech signals (2 from a male and 2 from a female speaker) using three distinct reverberation approaches:

- Artificial reverberation: This approach considered 6 distinct artificially generated RIRs, with the early reflections modeled by the image method [10], with a fixed source-microphone distance $d = 1.8$ m in a virtual room of dimensions length×width×height = 4 m × 3 m × 3 m. For $T_{60} = \{200, 300, 400\}$ ms, the late reverberation was emulated by a feedback-delay network [11], and for $T_{60} = \{500, 600, 700\}$ ms a modified version of Gardner's method [12, 13] was employed.
- Natural reverberation: In this case, the RIRs were obtained from the direct recordings of 4 different types

of rooms (booth, meeting, office, and lecture rooms) with several source-microphone distances for each room, as detailed in [14]. The 4 rooms have different wall dimensions and source-microphone distances making a total of different 17 RIRs. The average measured reverberation time for the 4 rooms are in the range of $\{120, 230, 430, 780\}$ ms.

- Real reverberation: In this case, the anechoic signals were directly played/recorded in 7 different rooms, and the corresponding RIRs were obtained from deconvolution. Each for the 7 rooms (booth, office1, lecture1, meeting1, office2, lecture2, meeting2) considered 4 source-microphone distances, $\{1, 2, 3, 4\}$ m, except for the smaller one (booth), where only 3 distances were employed, $\{0.5, 1, 1.5\}$ m, generating a total of 27 RIRs with average measured reverberation time in the range of $\{140, 390, 570, 650, 700, 890, 920\}$ ms.

In this Section the complete NBP database is employed to evaluate the performance of the original (with $L = 10$, $\mu = 3 \times 10^{-9}$ and $N_i = 500$) and modified (with $L = 50$, $\mu = 7 \times 10^{-9}$ and varying N_i) versions of the two-stage dereverberation algorithm.

Table 1 shows the Allen's score and T_{60} values (as estimated by the algorithm described in [15]) for the NBP processed signals using the two algorithm versions.

Table 1: *Mean performances of quality assessment measures (Allen's score (P) and T_{60}) for both algorithm versions.*

Quality Measures	Entire Database	Two-stage algorithm	
		Original	Improved
P	-2.95	-2.31	**-2.04**
T_{60} [ms]	517	335	**304**

Comparing the performances for the 200-signal database, one notices that the original and modified algorithm versions respectively achieved performance improvements of 22% and 30% for Allen's score and 35% and 40% for T_{60}. These results correspond to performance improvements for the modified algorithm in comparison to its original framework of approximately 12% for Allen's score and 9% for T_{60}. A t-test for this experiment indicated a confidence value of about 94.4% that the P and T_{60} mean results are significantly different for both algorithm versions.

For the entire NBP database, the average value of N_i was 21, representing an average computational-time reduction of 96%.

5. Conclusions

This paper analyzed the influence of several parameters for a two-stage enhancement algorithm for reverberant speech with respect to the perceived quality of the dereverberated signals. Three proposals of improvements in the original algorithm were analyzed, providing higher perceived quality at a cost of reduced computational complexity. The proposals consisted of analyzing the LP filter order, the adaptation stepsize μ and a convergence criterion for the inverse-filter modeling stage. The overall system performance was addressed for a 200-signal database of reverberated signals, leading to performance improvements of 12% and 9% in Allen's perceptual score and average T_{60}, at an average 96% reduction in the associated computational cost. Future work should include the use of a training database and different perceptual measures to assess the system performance.

6. Acknowledgements

The authors would like to thank Prof. M. Karjalainen, for providing the T_{60} estimation algorithm, and Prof. D. Wang, for providing the original two-stage algorithm for reverberant speech enhancement.

7. References

[1] R. Appel and J. Beerends, "On the Quality of Hearing Ones Own Voice," *J. Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, April 2002.

[2] M. Wu and D. Wang, "A Two-Stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 14, May 2006.

[3] J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," *Proc. IEEE Int. Workshop Acoustic Echo and Noise Control*, Paris, France, Sept. 2006.

[4] B. W. Gillespie, H. S. Malvar and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* Salt Lake, USA, May 2001.

[5] S. Haykin, *Adaptive filters theory,* 4th ed., Upper Saddle River, N.J.: Prentice-Hall, 2002.

[6] J. B. Allen, "Effects of small room reverberation on subjective preference," *J. Acoustic. Soc. Am.*, vol. 71, Apr. 1982.

[7] D. A. Berkley and J. B. Allen, "Normal Listening in Typical Rooms: The Physical and Psychophysical Correlates of Reverberation," *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., G.A. Studebaker and I. Hochberg eds., Allyn and Bacon, 1993.

[8] T. H. Falk and W.-Y. Chan, "A non-intrusive quality measure of dereverberated speech," *Proc. IEEE Int. Workshop Acoustic, Echo and Noise Control*, Seattle, USA, Sept. 2008.

[9] J. J. Jetz, "Critical distance measurement of rooms from the sound energy spectral response," *J. Acoustic. Soc. Am.*, vol. 65, pp. 1204–1211, May 1979.

[10] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustic. Soc. Am.* vol. 65, no. 4, pp. 943–950, Apr. 1979.

[11] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," *Proc. 90th Conv. Am. Engineering Soc.*, Preprint 3030, Feb. 1991.

[12] W. G. Gardner, *Reverberation Algorithms*, in *Applications of Digital Signal Processing*, Ed. Mark Kahrs and Karl-Heinz Brandenburg, Kluwer, New York:NY, pp. 85–131, Mar. 1998.

[13] A. A. de Lima, F. P. Freeland, P. A. A. Esquef, L. W. P. Biscainho, B. C. Bispo, R. A. de Jesus, S. L. Netto, R. Schafer, A. Said, B. Lee, and A. Kalker, "Reverberation assessment in audioband speech signals for telepresence systems," *Proc. Int. Conf. Signal Processing in Multimedia Applications*, Porto, Portugal, pp. 257–262, July 2008.

[14] M. Jeub, M. Schäfer, and P. Vary, "A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms," *Proc. 16th Int. Conf. on Digital Signal Processing*, Santorini, Greece, 2009.

[15] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy reponse measurements," *Proc. Conv. Audio Engineering Society*, Amsterdam, Netherlands, pp. 867–878, May 2001.