A SINGLE-MICROPHONE APPROACH FOR SPEECH SIGNAL DEREVERBERATION

Fernando S. Pacheco and Rui Seara

LINSE – Circuits and Signal Processing Laboratory
Department of Electrical Engineering
Federal University of Santa Catarina
88040-900 – Florianópolis – SC – Brazil
E-mails: {fernando, seara}@linse.ufsc.br

ABSTRACT

This paper presents a single-microphone speech dereverberation approach. The technique proposed comprises the room impulse response estimation followed by an inverse filtering. The estimation is carried out by identification of function zeros associated with the degradation system (room impulse response) in speech segments. A comparison between the cepstrum-based and proposed approaches is shown. Experimental results confirm the effectiveness and applicability of the new approach for dereverberation of mixed-phase impulse responses.

1. INTRODUCTION

In the transmission path between the source and the receiver, the speech signal (as any acoustic signal) is subject to modifications. This phenomenon affects the signal characteristics and it can be subjectively perceived as pleasant or not. In general, distortions can be grouped into two categories: additive and convolutional noise [1]. The latter is related to the room acoustic properties and/or the impulse response of the acquisition system [1]. In telephony applications, by using conventional telephones, the distance between the source (mouth) and the receiver (microphone) is small. In this way, distortions associated with the room environment are generally of low level. In contrast, in hands-free telephones the microphone is far from the speaker. Therefore, the captured signal contains the original signal as well as attenuated and delayed copies of this signal generated by reflections on the walls and other room surfaces. This signal degradation harms the intelligibility and comprehension of the spoken message. In speech recognition applications, system performance can be seriously affected by signal contamination. In these cases, dereverberation techniques are fundamental to reduce such degradation and increase speech intelligibility.

The most well-known dereverberation approaches have considered signal acquisition by microphone arrays. A classical technique that uses microphone arrays is the delay and sum approach and its variants [2]–[4]. In such an approach, microphones are placed in separate points within a room, and the speech signal travels different paths until reaching each microphone. Distance variations can be compensated through considering adequate delays in a way such that the sum of microphone signals reinforces the original signal and reduces the effect of acoustic reflections.

Speech dereverberation by using the signal captured by a single microphone and without previous knowledge about the room impulse response has been a challenging task. In [5], a single-microphone dereverberation approach is discussed, in which homomorphic filtering techniques are applied. In such an approach, it is assumed that the room impulse response and the original speech signal occupy separate regions in the cepstral domain. In this case, if complex cepstral components of the degraded signal, which are associated with the reverberation, present an impulsive structure, a cepstral filtering procedure (using a comb filter) can be considered for reducing (or even eliminating) the reverberation effect [5]. An alternative approach, also discussed in [5], is through the homomorphic filtering of the weighted complex cepstrum obtained from the contaminated signal. In this way, the components associated with the original speech signal (low quefrency) are isolated from those corresponding to the room impulse response (high quefrency). Cepstral analysis can also be used to estimate the room impulse response, as discussed in [6]. Satisfactory results related to its estimation are obtained for minimum-phase or mixed-phase responses which have a few zeros outside the unit circle in the z-plane [6]. In these research works, a major drawback from the cepstral approach is the mandatory condition that the original speech signal and room impulse response must occupy non-overlapping regions. In general, such an assumption is valid for minimum-phase impulse responses. However, for mixed-phase responses there are contributions from the room response in the low quefrency region. In real conditions, acoustic room responses have mixed-phase characteristic [1], restraining the use of dereverberation techniques based on cepstral analysis.

Another approach recently presented in the literature [7] takes advantage of the Bayesian framework for estimation of a source as a time-varying autoregressive (AR) process and the distortion by an all-pole filter. Although good results are obtained for gramophone recording restoration, the application in room dereverberation is not yet a consolidated approach.

This paper presents a new approach for estimating the room impulse response. The main contribution of the proposed technique is in the estimation of acoustic responses with mixed-phase characteristic. After the estimation of the room response, dereverberation using inverse response becomes a straightforward solution.

This paper is organized as follows. Section 2 presents the proposed speech dereverberation approach, mainly discussing the impulse response estimation technique. Experimental results, considering mixed-phase room responses, are presented in Section 3. Section 4 presents the conclusions and remarks of this research work.

This work was partially supported by CNPq, Brazil.

2. SPEECH DEREVERBERATION

The dereverberation approach used in this work consists of inverse filtering the distorted speech signal by the impulse response of the degradation system (room impulse response). This process is composed of three phases: (i) room impulse response estimation; (ii) determination of the inverse response; and (iii) signal filtering by the obtained inverse. Each of these phases is discussed in detail henceforth.

2.1 Room impulse response estimation

Speech signal distortion caused by reverberation can be modeled through a linear convolution operation in the sequence domain. Therefore, the modified speech signal $y(n)$, acquired by a single microphone, can be expressed as

$$y(n) = s(n) * h(n), \tag{1}$$

where $s(n)$ represents the original speech signal, $h(n)$ denotes the room impulse response, and "$*$" characterizes the linear convolution operation.

In the z-transform domain (1) is given by

$$Y(z) = S(z)H(z), \tag{2}$$

where $Y(z)$, $S(z)$, and $H(z)$ denote z-transforms of $y(n)$, $s(n)$ and $h(n)$, respectively.

Then, representing $S(z)$ and $H(z)$ in the factored form, we have

$$S(z) = A\prod_{k=1}^{M_1}(1-r_k z^{-1})\prod_{k=1}^{M_2}(1-s_k z^{-1})(1-s_k^* z^{-1}) \tag{3}$$

$$H(z) = B\prod_{k=1}^{M_3}(1-g_k z^{-1})\prod_{k=1}^{M_4}(1-h_k z^{-1})(1-h_k^* z^{-1}) \tag{4}$$

where A represents a gain, r_k and $\{s_k, s_k^*\}$ represent, respectively, M_1 real zeros and M_2 complex-conjugate pairs of zeros associated with $S(z)$. In function $H(z)$, B denotes a gain, g_k and $\{h_k, h_k^*\}$ represent, respectively, M_3 real zeros and M_4 complex-conjugate pairs of zeros.

Now, let us consider a speech signal partitioned into N segments. Each segment must be at least as long as the room impulse response length. Evaluating $Y(z)$ for each segment, we note that due to the time-varying nature of the speech signal represented by $S(z)$, it is unlikely to occur common zeros between these segments. On the other hand, assuming that the response $h(n)$ does not change considerably with time, zeros associated with $H(z)$ are kept in their fixed positions in the z-plane. Thus, from the zero constellation of $Y(z)$, evaluated for a certain number of segments, is possible to identify a fixed pattern (or with small change). Such a pattern is related to $H(z)$, which allows to determine an estimate of $h(n)$. Note that this model is valid if both speaker and microphone are spatially stationary, i.e., we are assuming that both (source and receiver) are not moving in the room.

Therefore, based on this principle, which is conceptually simple, the estimation steps of the room impulse response include: (i) segmentation of the reverberant signal; (ii) root finding for each segment; (iii) identification of zeros associated with the room impulse response; and (iv) unfactoring to obtain an estimate of the room impulse response.

In the segmentation process, considerations related to zeros of $Y(z)$ are still valid if some conditions are fulfilled. As mentioned previously, each segment should completely contain the response $h(n)$. In addition, to decrease the contamination of the current segment by the "tail" of the previous segment, the same strategy used in [6] is also applied here. Such an approach consists of using an exponential window for reducing the segmentation error. This window must start at a point after a silence period, overlapping the entire considered speech segment. Such a weighting window is defined by $w(n) = \gamma^n$, for $0 < \gamma < 1$.

Since segments are relatively long, high-order polynomials must be evaluated in the factorization process. In this condition, classical root-finder methods like Newton or other approaches based on eigenvalues of the matrix associated with the polynomial present serious convergence problems as well as a considerable computational burden. Thus, an interesting alternative is the use of the Lindsey-Fox root-finder [8], suitable for high-order polynomials, since such polynomials exhibit a zero constellation near the unit circle [8]. The central idea of this algorithm is the use of fast Fourier transform (FFT) for polynomial evaluation in concentric rings centered on the origin and with radius close to 1.

For the identification of the fixed-zero pattern along with $H(z)$, the proposed strategy is to represent the zeros of $Y(z)$ ($z_i = \alpha_i e^{j\theta_i}$) through a quantized singularity array, approach similar to [9]. This technique presents better results in terms of accuracy and speed over conventional clustering algorithms [9]. In our work, part of the z-plane is mapped into a nonlinear grid. This array is initially filled in with zero values. For each segment analyzed, those cells that match the location of each zero of the segment are incremented by 1. Thus, at the end of the evaluation process of N segments, the cells that represent the fixed zeros of $H(z)$ register a value N. In this way, the zero identification of $H(z)$ becomes simple and fast.

Some considerations related to the grid generation are important to improve the resolution of the identification procedure. This grid can be represented in either rectangular or polar coordinates. Since the zeros of $Y(z)$ are more concentrated around the unit circle, the magnitude–phase grid becomes a better alternative. The phase quantization is composed of 2^{B_p} locations, where B_p denotes the number of bits used to quantize the phase. In a similar way, the magnitude quantization presents 2^{B_m} positions, where B_m denotes the number of bits used to represent the magnitude. For responses composed of only real coefficients, it is sufficient to evaluate only the upper half z-plane (phase between 0 and π). For the problem in question, the phase distribution of zeros in this interval is considered uniform; thus, we use linear quantization for the phase. The magnitude of the zeros of $Y(z)$ has a distribution with concentration around unity. Then, a nonlinear quantization seems to be more efficient for the magnitude of the zeros. In this work, we adopt a shifted tangent function for such quantization. The value of the nonlinear quantized magnitude $\tilde{\alpha}_{i\mathrm{q}}$ is obtained by the following mapping relation:

$$\tilde{\alpha}_{i\mathrm{q}} = \frac{1}{\tan(\pi/D)}\left[\tan\left(\frac{\pi}{D}\right) + \tan\left(\frac{\pi}{D}\alpha_{i\mathrm{q}} - \frac{\pi}{D}\right)\right], \tag{5}$$

where D is a control parameter of the function, and $\alpha_{i\mathrm{q}}$ is the linearly quantized magnitude value α_i. Parameter D is adjusted to concentrate the grid around unity and to allow values between a minimum $\tilde{\alpha}_\mathrm{q}^{\min}$ and a maximum $\tilde{\alpha}_\mathrm{q}^{\max}$.

For practical reasons, zeros of the segments are split into two groups: zeros inside and outside the unit circle. For the former group, we use the procedure previously described. For the latter group, the procedure is applied to the zero reciprocals, allowing to consider for both groups $0 \leq \tilde{\alpha}_{iq} \leq 1$. Note that this strategy is only used for increasing the numerical accuracy in the identification stage. After the identification of zeros associated with $H(z)$, the zero reciprocals considered are remapped outside the unit circle, and an estimated response $\hat{h}(n)$ is obtained by unfactoring.

2.2 Inverse response determination and filtering

After the impulse response has been estimated, the inverse response must be determined for the deconvolution process of $y(n)$. Since, in general $\hat{h}(n)$ is of mixed phase, its inverse cannot be obtained directly, because it could lead to an unstable or noncausal system [10]. In this case, if some delay is tolerated, a least-squares technique can be used [11]. Another alternative is the decomposition of $\hat{h}(n)$ in minimum-phase and all-pass components by means of cepstral analysis [10], [12]. Using an iterative extraction of the minimum-phase component as presented in [10], problems associated with singularities close to the unit circle are overcome.

In the last stage, the estimated signal $\hat{s}(n)$ can be obtained by a convolution operation in the sequence domain between $y(n)$ and the inverse response $\hat{g}(n)$ or by a simple multiplication operation in the frequency domain. Thus, the latter reads

$$\hat{S}(e^{j\omega}) = Y(e^{j\omega})\hat{G}(e^{j\omega}), \tag{6}$$

where $\hat{S}(e^{j\omega})$, $Y(e^{j\omega})$, and $\hat{G}(e^{j\omega})$ are the Fourier transforms of $\hat{s}(n)$, $y(n)$ and $\hat{g}(n)$, respectively.

3. EXPERIMENTAL RESULTS

The proposed approach is applied and compared with cepstral processing [6] (which also uses only one microphone) in terms of dereverberated speech improvement.

A room impulse response is generated by using a computer implementation of the image method [13]. The room size is assumed to be 3.54 m (length) $\times$ 2.62 m (width) $\times$ 2.97 m (height). In the simulation scenario, the speaker is located at position (1.9, 1.49, 0.86) (m) and the microphone is positioned at (0.45, 1.77, 1.74) (m). The walls of the room have the same reflection coefficient of 0.9, while the coefficient for the floor and ceiling is 0.8. The sampling rate used is 8 kHz. Note that the generated response has 158 zeros outside the unit circle. This response is shown in Fig. 1.

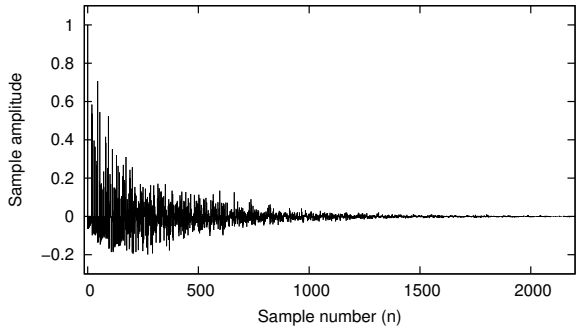

Figure 1: Room impulse response.

Speech segments consisting of seven isolated digits in Portuguese, with a total duration of 5 s, sampled at 8 kHz, are used as the original signal $s(n)$. Fig. 2 shows one of these speech segments, corresponding to the number eight ("oito" in Portuguese). Note that such a number of segments is necessary for the cepstral processing, in which a averaging operation must be accomplished. For our approach, only two or three segments are sufficient. The reverberant speech signal $y(n)$ is generated by means of a linear convolution between $s(n)$ and the impulse response. Fig. 3 now presents the same speech segment convolved with the room impulse response.

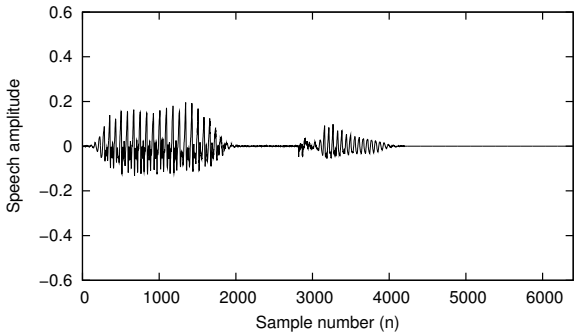

Figure 2: Segment of original signal.

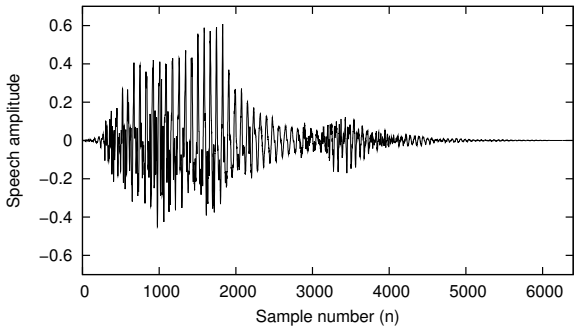

Figure 3: Segment of reverberant signal.

In this work, segmentation points are manually selected (for both approaches), although an automatic segmentation could be easily carried out.

For each segment, zeros associated with the signals are determined by using the Lindsey-Fox root-finder [8]. Zeros of each segment are mapped into the quantization grid. In the considered example, the parameters used are $B_{\mathrm{p}} = 12$ bits, $B_{\mathrm{m}} = 10$ bits, $D = 2.05$, $\tilde{\alpha}_{\mathrm{q}}^{\min} = 0$ e $\tilde{\alpha}_{\mathrm{q}}^{\max} = 1$. Selecting the cells which register the number of used segments we obtain the zeros associated with $\hat{h}(n)$. After unfactoring, $\hat{h}(n)$ is obtained. From the estimated response, the inverse response is obtained using the procedure described in [10] and the dereverberated signal is determined by the convolution of $y(n)$ with $\hat{g}(n)$. Fig. 4 presents the dereverberated signal using the cepstrum-based processing and Fig. 5 shows the result of the proposed dereverberation approach.

We also use an objective measure to assess the performance of the proposed approach and compare it to the cepstrum-based processing. Signal-to-noise ratio (SNR) can

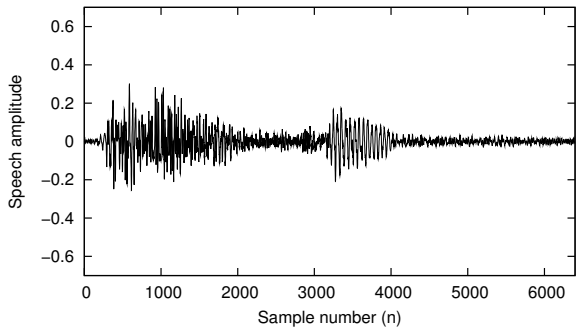

Figure 4: Dereverberated signal using cepstral processing.

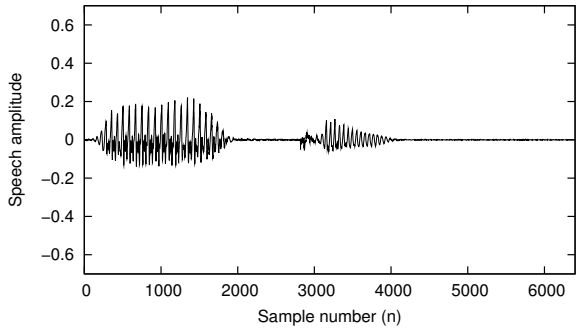

Figure 5: Dereverberated signal using the proposed approach.

be interpreted as a direct-to-reverberant signal component ratio (DRR) [4], which is defined as

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{i=0}^{L-1}[s^2(i)]}{\sum_{i=0}^{L-1}[\hat{s}(i)-s(i)]^2}, \qquad (7)$$

where $s(n)$ and $\hat{s}(n)$ represent the original and processed signals, respectively, L denotes the length of the signal and i characterizes the sample index. The DRRs for the reverberant speech, processed speech using cepstrum and the proposed approach are, respectively, -9.34, -3.5, and 17.4 dB. The proposed approach represents an effective improvement over the cepstrum-based processing. Note that the cepstral approach could attain a better result in a scenario involving a minimum-phase response, but this does not correspond to a practical situation.

4. CONCLUSIONS

In this paper a new approach for room impulse response estimation is proposed. Experimental results show the applicability of our technique in estimating mixed-phase responses for speech dereverberation applications by using a single microphone. The proposed approach provides an interesting solution for speech recognition applications in which the speech signal is contaminated by reverberation. In further work, we intend to evaluate the performance in real room responses considering an automatic speech segmentation strategy.

5. ACKNOWLEDGEMENT

The authors are grateful to Mr. James W. Fox for making available the source code of the Lindsey-Fox algorithm for factoring very-high-degree polynomials.

REFERENCES

[1] S. Subramanian, A. P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 5, pp. 392–396, Sept. 1996.

[2] J. B. Allen, D. A. Berkley, and J. Blauert, "A multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, pp. 912–915, Oct. 1977.

[3] K. Eneman, J. Duchateau, M. Moonen, D. V. Compernolle, and H. V. Hamme, "Assessment of dereverberation algorithms for large vocabulary speech recognition systems," in *Proc. Europ. Conf. Speech Commun. Technol. (EUROSPEECH'03)*, Geneva, Sept. 2003, pp. 1–4.

[4] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Multimicrophone speech dereverberation via spatio-temporal averaging," in *Proc. Europ. Signal Processing Conf. (EUSIPCO'04)*, Vienna, Sept. 2004, pp. 809–812.

[5] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham Jr., "Nonlinear filtering of multiplied and convolved signals," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, no. 3, pp. 437–466, Sept. 1968.

[6] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'91)*, vol. II, Toronto, May 14–17, 1991, pp. 977–980.

[7] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 476–488, Sept. 2003.

[8] G. A. Sitton, C. S. Burrus, J. W. Fox, and S. Treitel, "Factoring very-high-degree polynomials," *IEEE Signal Processing Mag.*, vol. 20, no. 6, pp. 27–42, Nov. 2003.

[9] B. Theobald, S. Cox, G. Cawley, and B. Milner, "Fast method of channel equalisation for speech signals and its implementation on a DSP," *Electron. Lett.*, vol. 35, no. 16, pp. 1309–1311, Aug. 1999.

[10] B. D. Radlović and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, pp. 728–737, Nov. 2000.

[11] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound Vibr.*, vol. 102, no. 2, pp. 217–228, Sept. 1985.

[12] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Englewood Cliffs: Prentice Hall, 1989.

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