BINAURAL EXTENSION AND PERFORMANCE OF SINGLE-CHANNEL SPECTRAL SUBTRACTION DEREVERBERATION ALGORITHMS

Alexandros Tsilfidis, Eleftheria Georganti, John Mourjopoulos

Audio and Acoustic Technology Group
Wire Communications Laboratory
Department of Electrical and Computer Engineering
University of Patras, 26504, Patras, Greece

ABSTRACT

Single-channel spectral subtraction algorithms are commonly used to suppress late reverberation. A binaural extension of such methods, apart from suppressing reverberation without introducing processing artifacts, should also preserve the signal's binaural localization cues. Here, three state-of-the-art spectral subtraction dereverberation algorithms are extended into a binaural context utilizing three alternative bilateral gain adaptation schemes and are compared to an extension derived from a Delay and Sum Beamformer. Objective results for several experimental conditions reveal the most prominent binaural extensions.

Index Terms— Binaural dereverberation, spectral subtraction, speech enhancement

1. INTRODUCTION

Room reverberation degrades speech quality and intelligibility and also reduces the Automatic Speech Recognition(ASR) performance. Hence, blind or semi-blind dereverberation methods have been developed, utilizing single or multiple input channels. Dereverberation is also important for binaural applications in the context of digital hearing aids, binaural telephony and hands free devices (e.g. [1, 2, 3]). However, adapting single or multichannel techniques for binaural processing is not trivial. Apart from the challenging task of reducing reverberation without introducing audible artifacts, binaural dereverberation methods should also at least preserve the Interaural Time Difference (ITD) and Interaural Level Difference (ILD) cues as it has been shown that bilateral signal processing affects the source localization [3].

Most dereverberation techniques tackle separately the coloration produced by the early reflections and the overlap-masking produced by the decaying reverberant tails; the latter being mainly responsible for the adverse effects in speech communication and ASR systems. Recently, Jeub et al.[3] proposed a two-stage dereverberation algorithm that explicitly preserves binaural cues. The coloration is reduced with a dual-channel Wiener-filter while late reverberation is suppressed through spectral subtraction employing a binaural version of a well-known technique [4].

The research activities that led to these results were co-financed by Hellenic Funds and by the European Regional Development Fund (ERDF) under the Hellenic National Strategic Reference Framework (NSRF) 2007-2013 according to Contract no. MICRO2-38/E-II-A of the project "MEMSENSE" within the Programme "Hellenic Technology Clusters in Microelectronics - Phase-2 Aid Measure".

The present work examines and evaluates efficient techniques to adapt single-channel spectral subtraction dereveberation algorithms to a binaural context. Hence, binaural extensions based on the Delay and Sum Beamformer (as proposed in [3]) are evaluated for three state-of-the-art spectral subtraction methods [4, 5, 6]. In addition, a generalized approach based on the adaptation of the spectral gains derived by bilateral processing is presented and three possible gain adaptation strategies are investigated.

2. SINGLE-CHANNEL SUPPRESSION BASED ON SPECTRAL SUBTRACTION

The basic principle of single-channel spectral subtraction dereverberation [4, 5, 6] is to estimate the short time spectrum of the clean signal $S_e(\omega, j)$ by subtracting an estimation of the short time spectrum of late reverberation $R(\omega, j)$ from the short time spectrum of the reverberant signal $Y(\omega, j)$:

$$S_e(\omega, j) = Y(\omega, j) - R(\omega, j) \tag{1}$$

where ω and j are the frequency bin and time index respectively. Following an alternative formulation, the estimation of the short time spectrum of the clean signal can be derived by applying appropriate weighting gains $G(\omega, j)$ in the short time spectrum of the reverberant signal i.e.:

$$S_e(\omega, j) = G(\omega, j)Y(\omega, j) \tag{2}$$

where

$$G(\omega, j) = \frac{Y(\omega, j) - R(\omega, j)}{Y(\omega, j)} \tag{3}$$

Therefore, the problem is deduced in an estimation of the late reverberation short time spectrum.

Lebart et al. [4] proposed a method based on the Room Impulse Response (RIR) modeling (LB), where the RIR is modelled as a discrete non-stationary stochastic process. The modeling is valid when the direct energy of the RIR is smaller than the energy of the reflections [7]. The short time spectral magnitude of the reverberation is estimated as:

$$|R(\omega, j)| = \frac{1}{\sqrt{|SNR_{pri}(\omega, j)| + 1}} |Y(\omega, j)| \tag{4}$$

where $|SNR_{pri}(\omega, j)|$ is the a priori Signal to Noise Ratio that can be approximated by a moving average of the a posteriori Signal to Noise Ratio $|SNR_{post}(\omega, j)|$ in each frame:

$$|SNR_{pri}(\omega, j)| = \beta |SNR_{pri}(\omega, j-1)| + (1-\beta) max(0, |SNR_{post}(\omega, j)|) \tag{5}$$

The method proposed by Wu and Wang [5] (WW) is motivated by the observation that the smearing effect of late reflections produces the smoothing of the signal spectrum in the time domain. Hence, the late reverberation power spectrum is considered a smoothed and shifted version of the power spectrum of the reverberant speech:

$$|R(\omega,j)|^2 = \gamma w(j-\rho) * |Y(\omega,j)|^2 \quad (6)$$

where ρ is a frame delay, γ a scaling factor. The term $w(j)$ represents an assymetrical smoothing function given by the Rayleigh distribution:

$$w(j) = \begin{cases} \frac{j+\alpha}{\alpha^2} \exp\left(\frac{-(j+\alpha)^2}{2\alpha^2}\right) & \text{if } j < -\alpha \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where α represents a constant number of frames. Both in [4] and [5], the phase of the reverberant speech is combined with the estimated clean signal's spectrum and overlap add is used to extract the time domain estimation.

Alternatively, Furuya and Kataoka [6] proposed a method (FK) where the short time power spectrum of late reverberation in each frame can be estimated as the sum of filtered versions of the previous frames of the reverberant signal's short time power spectrum:

$$|R(\omega,j)|^2 = \sum_{l=1}^{K} |a_l(\omega,j)|^2 |Y(\omega,j-l)|^2 \quad (8)$$

where the coefficients of late reverberation $a_l(\omega,j)$ are derived from:

$$a_l(\omega,j) = E\left\{ \frac{Y(\omega,j)Y^*(\omega,j-l)}{|Y(\omega,j-l)|^2} \right\} \quad (9)$$

Then an estimation of the clean signal in the time domain can be derived through overlap add from the short-time spectrum of the dereverberated signal $S_e(\omega,j)$:

$$S_e(\omega,j) = \left\{ \frac{|Y(\omega,j)|^2 - |R(\omega,j)|^2}{|Y(\omega,j)|^2} \right\} Y(\omega,j) \quad (10)$$

3. BINAURAL DEREVERBERATION PROCESSING

An effective approach for extending the LB method to a binaural context is to derive a reference signal using a Delay and Sum Beamformer (DSB) [3] where the time delays are estimated utilizing a method based on the generalized cross-correlation with phase transform as proposed in [8]. The reference signal is calculated as the average of the time aligned left and right reverberant signals. Using the reference, appropriate weighting gains are derived following Equations 2, 3 and 4, and identical processing is applied to both left and right channel. Here, the DSB approach is also implemented for both the WW and FK methods in order to evaluate the efficiency of different late reverberation estimation techniques in a binaural scenario.

In addition, an alternate binaural adaptation scheme is proposed. In binaural applications, the time delay between the left and right channels of the speech signal is limited by the width of the human head. Therefore, it can be assumed shorter than the length of a typical analysis window used in spectral subtraction techniques and hence the time alignment stage is omitted. Then, each algorithm is implemented independently for the left and right ear channel signals resulting to the corresponding weighting gains $G_l(\omega,j)$ and $G_r(\omega,j)$. These gains are combined and different adaptation strategies are investigated for each algorithm:

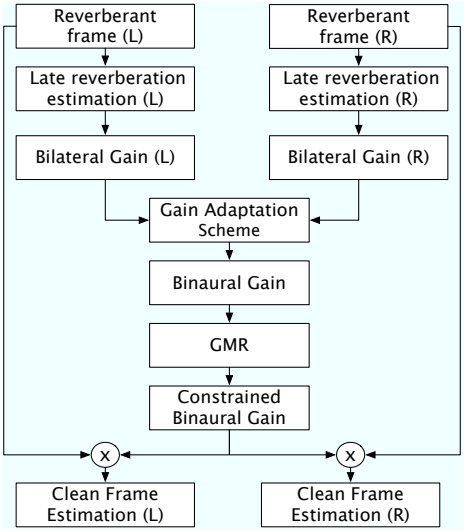

Fig. 1. Frame-level processing of the proposed implementation

(i) The final gain is derived as the maximum of the left and right channel weighting gains:

$$G(\omega,j) = max(G_l(\omega,j), G_r(\omega,j)) \quad (11)$$

This approach (maxGain) achieves moderate late reverberation suppression, but it is also less likely to produce overestimation artfacts.
(ii) The final gain is derived as the average of the left and right channel weighting gains:

$$G(\omega,j) = \frac{(G_l(\omega,j) + G_r(\omega,j))}{2} \quad (12)$$

This gain adaptation strategy (avgGain) compensates equally for the contribution of the left and right channels.

(iii) The final gain is derived as the minimum of the left and right channel weighting gains:

$$G(\omega,j) = min(G_l(\omega,j), G_r(\omega,j)) \quad (13)$$

The above adaptation technique (minGain) results to maximum reverberation attenuation but the final estimation may be susceptible to overestimation artifacts.

When late reverberation is overestimated, spectral subtraction methods tend to generate musical noise artifacts. Many methods have been proposed in order to compensate for this effect (e.g. [9, 3]). Here a low-complexity approach based on Gain Magnitude Regularization (GMR) is used. An overestimation of the late reverberation is less likely to happen in high SRR spectral regions such as signal steady states [9] contrary to low SRR regions. Therefore a low SRR detector is employed [3] and a GMR technique is introduced in order to constrain only the low gain parts. Hence, the new constrained gain $G'(\omega,j)$ is derived as:

$$G'(\omega,j) = \begin{cases} \frac{G(\omega,j)-\theta}{r} + \theta & \text{when } \zeta < \zeta_{th} \cap\ G(\omega,j) < \theta \\ G(\omega,j) & \text{otherwise} \end{cases} \quad (14)$$

Table 1. Parameter values for the employed methods

Parameter	LB	WW	FK
Total Frame Length	1024	1024	2048
Zero padding	512	128	128
Frame Overlap	0.125	0.25	0.25

and

$$\zeta = \frac{\sum_{\omega=1}^{\Omega} G(\omega, j)|Y(\omega, j)|^2}{\sum_{\omega=1}^{\Omega} |Y(\omega, j)|^2} \quad (15)$$

where θ is the threshold for applying the gain constraints, r is the regularization ratio, ζ is the power ratio between the enhanced and the reference signal, ζ_{th} the threshold of the low SRR detector and Ω is the frame size.

When $\zeta \approx 1$ the power of the dereverberated (i.e. the "clean") frame is comparable to the power of the reverberant frame and hence a high SRR frame is assumed. On the other hand, when $\zeta \approx 0$ it is assumed that reverberation has been significantly suppressed and therefore the original frame had a low SRR value [3]. Moreover, the parameter r controls the regularization rate while the parameter θ discriminates the low from the high gain bins and the GMR is applied only on the low gain bins[1]. Fig. 1 presents a block diagram of the proposed framework.

4. TESTS AND RESULTS

Eight anechoic phrases uttered by both male and female speakers of the TIMIT database were convolved with real Binaural RIRs (BRIRs). Four BRIRs measured in a Stairway Hall ($RT_{60} = 0.69$ sec) at a source-receiver distance of $3m$ and azimuth angles of $0, 30, 60$ and 90° were chosen from the Aachen database [3]. In addition three BRIRs measured in a Cafeteria ($RT_{60} = 1.29$ sec) at source-receiver distances of $1.18,\ 1$ and $1.62\ m$ and azimuth angles of approximately $-30, 0$ and 90° were chosen from the Oldenburg database [10]. The speech signals and the BRIRs were sampled at 16 kHz with a 16 bit resolution and the authors made unofficial tests to select optimal values for the analysis parameters (see Table 1). The θ and ζ_{th} values of the GMR step were set at 0.15, the regularization ratio r was 4 and the RT_{60} was calculated from the impulse responses. All parameter values that are not detailed here were set according to the values proposed by the authors of the original works. In addition, for the FK and LB techniques, two additional relaxation criteria were imposed [9] as they were previously found by the authors to have advantageous effects on the performance. The WW and FK methods assume that an inverse-filtering stage precedes the spectral subtraction implementation. However here, the implementation of a 1/3 octave minimum-phase inverse filtering was not found to notably alter the relative improvement achieved by the tested methods. Therefore, a generalized case where the spectral subtraction is applied directly to the reverberant signals is presented.

The average segmental Signal to Reverberation Ratio (SRR) differences when compared to the corresponding reverberant signals for (a) Stairway Hall and (b) Cafeteria are presented in Fig. 2. The

[1]Audio demos illustrating the effect of the GMR and the three gain adaptation schemes can be found at `http://www.wcl.ece.upatras.gr/audiogroup/tools/derev.html`

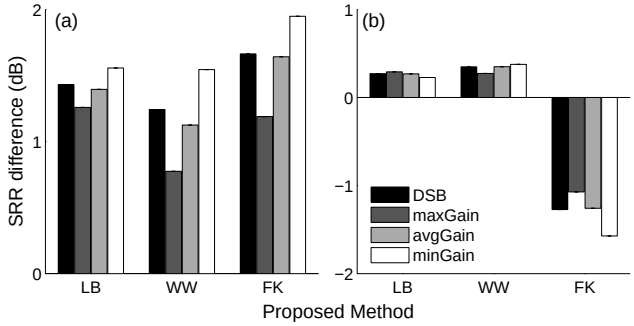

Fig. 2. SRR difference for the three tested methods (LB, WW, FK) for (a) Stairway Hall , (b) Cafeteria (DSB: Delay and Sum Beamformer, maxGain: maximum bilateral gain adaptation, avgGain: average bilateral gain adaptation, minGain: minimum bilateral gain adaptation).

SRR measure evaluates the suppression intensity and it is the equivalent of SNR when reverberation is considered as additive noise. For the case of the Stairway Hall, all binaural extension strategies for all three methods achieve a significant SRR improvement. As expected, the minGain technique achieved substantial reverberation suppression and therefore resulted to a greater SRR. On the other hand, less reverberation was suppressed utilizing the maxGain technique. The DSB and the avgGain adaptation techniques produce similar results as in principle they both take into account equally the left and right channel. The FK method seems to suppress more reverberation than the other two tested methods. The SRR differences presented in Fig. 2(b) for the larger enclosure (Cafeteria) are significantly smaller. In such rooms, dereverberation becomes a very challenging problem and most algorithms introduce artifacts due to late reverberation overestimation errors. Hence, it can be seen that both LB and WW approaches achieve a small SRR improvement while the FK method reduces the SRR in all cases.

A further evaluation of the produced signals is made through the Perceptual Speech Quality measure (PESQ) variation [11], compared to the reverberant signals. PESQ implements a perceptual model in order to assess the quality of a processed speech signal and rates it according to the five grade Mean Opinion Score (MOS) scale. The results are presented in Table 2 (bold values denote optimum performance). For the case of the Stairway Hall the bigger PESQ improvement is achieved utilizing the WW method with the minGain adaptation technique. The same technique seems to be also the optimal choice when used in conjunction with the LB method. It can be assumed that in a scenario where bilateral late reverberation estimations are successful this technique presents superior performance. However, it is not beneficial when used with the FK method where probably the bilateral processing resulted to inferior results. The FK method produces better results when used with the avgGain technique. In general, the WW method shows a significant PESQ increment for all tested adaptation techniques. For the Cafeteria, the LB method produces a relatively stable PESQ improvement independent of the employed binaural extension. On the other hand, better results are derived with the WW method for all binaural adaptation techniques; the best results achieved with the avgGain approach. The FK method seems to produce processing artifacts despite the utilized binaural adaptation scheme and decreases the PESQ values in every case. Note that all the above results were actually confirmed after several informal listening tests performed by the authors.

Table 2. PESQ improvement for various cases

Stairway Hall				
Method	DSB	maxGain	avgGain	minGain
LB	0.153	0.142	0.147	0.158
WW	0.206	0.160	0.208	**0.258**
FK	0.160	0.180	0.186	−0.029
Cafeteria				
Method	DSB	maxGain	avgGain	minGain
LB	0.133	0.136	0.135	0.133
WW	0.205	0.208	**0.216**	0.198
FK	−0.235	−0.141	−0.228	−0.428

Finally, note that the DSB implementation has the advantage of lower computational complexity as it involves single-channel processing for the estimation of the weighting gain functions. On the other hand, the proposed gain adaptation techniques involve bilateral processing, but do not necessitate the initial time delay estimation.

Spectral-subtraction dereverberation techniques were often proposed in order to improve speech intelligibility in teleconference setups and to compensate for the ASR deterioration. In such cases, it is usual to assume speech reproduction in acoustically treated rooms and the proposed dereverberation methods achieve better results. However, in real-life scenarios (e.g. binaural dereverberation for hearing aids) no acoustically optimized enclosures can be assumed. In such cases perceptually-motivated algorithms may be more appropriate (e.g. [12]). However, the binaural extension of such algorithms is challenging as it should take into account many aspects of the binaural hearing mechanism.

5. CONCLUSION

Different binaural implementation strategies for single-channel spectral subtraction dereverberation algorithms were presented. The performance of a previously proposed approach based on a Delay and Sum Beamformer was compared with three new schemes adapting the gains derived from bilateral processing. All techniques are implemented in three state-of-the-art spectral subtraction dereverberation methods. The results show that best performance in low reverberation environments is achieved when using the minGain or the avgGain technique while in strongly reverberant conditions the maxGain or the avgGain implementation achieve better results. Finally the method proposed by Wu and Wang [5] was found to be more robust for extension into a binaural context.

6. REFERENCES

[1] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Commun.*, vol. 39, pp. 111–138, 2003.

[2] H.W. Lollmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP J. Adv. Sig. Pr.*, vol. 2009, pp. 1–9, 2009.

[3] M. Jeub, M Schafer, T Esch, and P Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 18, pp. 1732–1745, 2010.

[4] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acust. Acust.*, vol. 87, pp. 359–366, 2001.

[5] M. Wu and D Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 14, pp. 774–784, 2006.

[6] K. Furuya and A Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, pp. 1579–1571, 2007.

[7] E.A.P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. thesis, Technische Univ. Eindhoven, 2007.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-24, pp. 320–327, 1976.

[9] A. Tsilfidis and J. Mourjopoulos, "Signal-dependent constraints for perceptually motivated suppression of late reverberation," *Signal Process.*, vol. 90, pp. 959–965, 2010.

[10] H. Kayser, S.D. Ewert, J. Anemuller, T. Rohdenburg, V. Hohhmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Adv. Sig. Pr.*, vol. 2009, pp. 1–10, 2009.

[11] ITU-R P.862, "Perceptual evaluation of speech quality (pesq), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," 2000.

[12] A. Tsilfidis and J. Mourjopoulos, "Blind single-channel suppression of late reverberation based on perceptual reverberation modeling," *J. Acoust. Soc. Amer.*, vol. 129(2), 2011.