A NEW CASCADED SPECTRAL SUBTRACTION APPROACH FOR BINAURAL SPEECH DEREVERBERATION AND ITS APPLICATION IN SOURCE SEPARATION

Muhammad Salman Khan, Syed Mohsen Naqvi, and Jonathon Chambers

Advanced Signal Processing Group, School of Electronic, Electrical and Systems Engineering,
Loughborough University, Leicestershire, LE11 3TU, UK.
Email: {m.s.khan2, s.m.r.naqvi, j.a.chambers}@lboro.ac.uk

ABSTRACT

In this work we propose a new binaural spectral subtraction method for the suppression of late reverberation. The proposed approach is a cascade of three stages. The first two stages exploit distinct observations to model and suppress the late reverberation by deriving a gain function. The musical noise artifacts generated due to the processing at each stage are compensated by smoothing the spectral magnitudes of the weighting gains. The third stage linearly combines the gains obtained from the first two stages and further enhances the binaural signals. The binaural gains, obtained by independently processing the left and right channel signals are combined using a new method. Experiments on real data are performed in two contexts: dereverberation-only and joint dereverberation and source separation. Objective results verify the suitability of the proposed cascaded approach in both the contexts.

Index Terms— Speech enhancement, binaural dereverberation, spectral subtraction, source separation

1. INTRODUCTION

Room reverberation, produced by multiple reflections of the sound on wall surfaces and objects in an enclosure, remains a challenge for many signal processing applications, such as automatic speech recognition (ASR), hearing aids and hands-free telephony. Specifically, the late reflections of the room impulse response (RIR) cause spreading of the speech spectra and degrade the quality of speech and the intelligibility [1]. The objective of dereverberation algorithms is to suppress the effects of reverberation while minimally distorting the speech structure.

Monaural dereverberation algorithms based on spectral subtraction, e.g. [1, 2], have been proposed to suppress the effects of late reflections. Effective extension of the monaural methods to the binaural context is important as this would enable their utilization in multiple applications. Such extensions must produce minimal musical noise and also preserve the binaural cues i.e. interaural time difference (ITD) and the interaural level difference (ILD) [3, 4].

In this paper, we investigate a novel cascaded binaural spectral subtraction method and an alternate binaural gain formation scheme. The cascading of non-linear processors has previously been reported to be useful for radio-frequency interference suppression [5], [6]. Our proposed cascaded approach utilizes the models of two state-of-the-art monaural methods [1] and [2]. Each of the three concatenated dereverberation blocks enhances the output of the previous stage by suppressing the effects of late reverberation. The weighting gains of the left and right channels, in the first two stages, are formed using a new linear convex combination. This provides additional flexibility for use at different levels of reverberation. The gain functions obtained in stage 1 and 2 are combined to form a new gain which is applied to the reverberant mixture to give the final dereverberated signal. Computational complexity is not addressed in this work, rather proof of concept is the focus. The proposed dereverberation method is also used in a unified algorithm for the purpose of joint dereverberation and source separation. The source separation algorithm is based on the models of ILD and ITD. The objective of such a unification is to emphasize the advantage of a robust dereverberation method as a pre-processing stage to a source separation system used in a highly reverberant context. Extensive experiments are carried out using real datasets to verify the advantage of the proposed scheme. In Section 2 we explain the monaural spectral subtraction methods and in Section 3 the proposed binaural dereverberation scheme. Section 4 discusses the experiments and results with the relation to prior work and conclusion following next.

2. MONAURAL SPECTRAL SUBTRACTION BASED SPEECH DEREVERBERATION

A general expression for spectral subtraction based dereverberation techniques in the short-time Fourier transform

Thanks to UET, Peshawar and the Higher Education Commission (HEC) of Pakistan for funding M. S. Khan, and the Engineering and Physical Sciences Research Council (EPSRC) of the UK (Grant number EP/H049665/1).

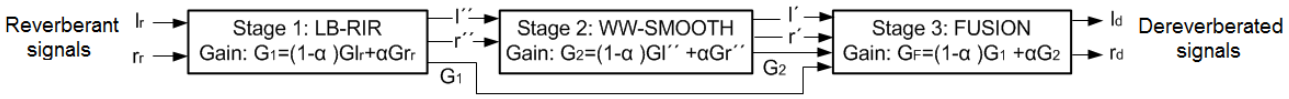

Fig. 1. The proposed cascaded approach for dereverberation. Input/output of each processing stage is in the time-domain.

(STFT) domain is written as:

$$X_{cln}(\omega,t) = X_{rev}(\omega,t) - X_{rev_{late}}(\omega,t) \quad (1)$$

where $X_{rev_{late}}(\omega,t)$ is the late reverberant component which is subtracted from $X_{rev}(\omega,t)$, the observed reverberant signal to give the clean signal $X_{cln}(\omega,t)$ at frequency index ω and time frame t. This process can also be expressed as

$$X_{cln}(\omega,t) = G(\omega,t)X_{rev}(\omega,t) \quad (2)$$

where $G(\omega,t)$ is a gain function applied to the observed reverberant signal $X_{rev}(\omega,t)$ to obtain the dereverberated clean signal $X_{cln}(\omega,t)$. Thus, the objective of the dereverberation schemes is to derive the gain by estimating the late reverberant component.

In [2], which we refer to as LB-RIR (the acronym is derived from the authors' names, Lebart et al., and their technique which is based on RIR modeling), the authors proposed to model statistically the RIR in order to subtract spectrally the late reverberant component. They exploit the observation of the smearing of the energy of the speech signal into reverberation tails due to overlap-masking. The model assumes that the direct-to-reverberant (DRR) ratio is low. The weighting gain is calculated as

$$G(\omega,t) = 1 - \frac{1}{\sqrt{SIR_{post}(\omega,t)+1}} \quad (3)$$

where $SIR_{post}(\omega,t) = \frac{|X_{rev}(\omega,t)|^2}{\sigma^2_{X_{rev_{late}}}(\omega,t)}$ is the *a posteriori* signal-to-interference ratio (SIR). Here $\sigma^2_{X_{rev_{late}}}(\omega,t)$ is the variance of the late reverberant speech component estimated as

$$\sigma^2_{X_{rev_{late}}}(\omega,t) = \exp(-2\kappa T_l) \,.\, \sigma^2_{X_{rev}}(\omega, t-n_{late}) \quad (4)$$

where $\kappa = \frac{3ln(10)}{RT60}$, T_l indicates the time from which the late reverberation starts, n_{late} is the number of samples related to T_l, $RT60$ indicates the reverberation time, and $\sigma^2_{X_{rev}}$ is the variance of the reverberant mixture computed by recursive averaging [3]

$$\sigma^2_{X_{rev}}(\omega,t) = \delta \,.\, \sigma^2_{X_{rev}}(\omega,t-1) + (1-\delta) \,.\, |X_{rev}(\omega,t)|^2 \quad (5)$$

where $\delta \in [0,1]$ is the smoothing factor.

The work presented in [1], which we term WW-SMOOTH (the acronym is derived from the authors' names, Wu and Wang, and their method which is based on smoothing of the signal spectrum), is motivated by the observation that the spreading due to the late reverberation causes smoothing of the signal spectrum in the time domain. Thus, the power of the late reverberant component is estimated as the smoothed and shifted version of the power of the reverberant speech in the TF domain

$$|X_{rev_{late}}(\omega,t)|^2 = \gamma\varpi(t-\rho) * |X_{rev}(\omega,t)|^2 \quad (6)$$

where $*$ indicates the convolution operation, γ is a scaling factor, and ρ is the shift delay. The term $\varpi(t)$ is a smoothing function given as the Rayleigh distribution:

$$\varpi(t) = \begin{cases} \frac{t-a}{a^2}\exp(\frac{-(t-a)^2}{2a^2}), & if\ t > -a \\ 0, & otherwise \end{cases}$$

where a indicates the number of frames and needs to be smaller than ρ. Binaural extension of the monaural dereverberation algorithms and the proposed scheme are discussed next.

3. PROPOSED BINAURAL SPEECH DEREVERBERATION

In [3] the monaural scheme LB-RIR is extended to a binaural form and a delay-and-sum beamformer is used to generate a reference signal. The time-aligned left and right reverberant signals are averaged to obtain the reference signal. The reference signal is then processed to generate a single gain mask using Eq. (3). Alternatively, in [4] the left and the right reverberant mixtures are separately processed to yield two gains. The two gains are then combined, e.g. by taking the minimum, maximum or average, and applied to both the channels.

We propose a new approach to dereverberate the reverberant speech signals and adopt a different scheme for combining the bilateral weighting gains. The entire dereverberation process is a combination of three cascaded stages. Each stage takes a binaural input and gives binaural outputs in the time-domain. The algorithm diagram is given in Fig. 1. The enhancement of each stage is cumulative. With a cascade of these non-linear processors, a higher overall enhancement is achievable which may not be possible by each stage individually, or by repeatedly cascading the same block.

The time-domain left and right channel reverberant signals are input to the first stage where they are independently processed using the LB-RIR scheme. The left and right gains are then linearly combined using the equation

$$G_{new} = (1-\alpha)G_a + \alpha G_b \quad (7)$$

where $\alpha \in [0,1]$, which is chosen empirically. Stage 1 outputs the time-domain left and right enhanced signals that are fed to the second stage where the method of WW-SMOOTH is applied to obtain the bilateral gains. The gains are combined similar to as in the first stage and are applied to the enhanced signals from stage 1. Fig. 2 depicts the processing of stages 1 and 2. The enhanced signals from stage 2 are forwarded to stage 3. The weighting gains from stage 1 and stage 2 are linearly fused to form a combined gain. The fused gain is used to further suppress the late reverberant components from the left and right channel signals and give the final dereverberated signals.

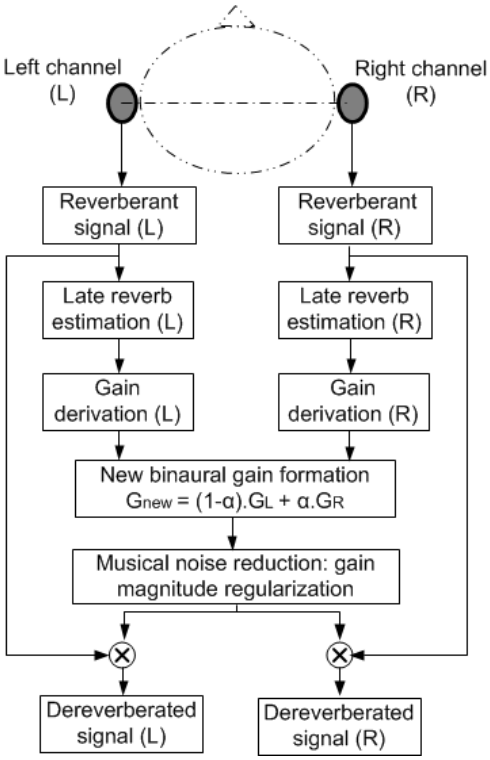

Fig. 2. Processing overview for the first two stages with the bilateral signal processing and new gain derivation.

In stages 1 and 2, smoothing of the weighting gains is performed in order to reduce the effects of the musical tones that are commonly produced due to overestimation of the late reverberation in speech-less or low SIR regions. The power ratio of the enhanced signal and the reverberant signal [7],

$$\zeta(t) = \frac{\sum_{\omega=1}^{W} G_{new}(\omega,t).|Y(\omega,t)|^2}{\sum_{\omega=1}^{W} |Y(\omega,t)|^2} \quad (8)$$

is computed to indicate whether the SIR of a time frame is low or high. If $\zeta(t)$ is approximately unity, the SIR of that frame is assumed to be high, and if $\zeta(t)$ is nearly zero, the SIR is supposed to be low. A moving average window is then applied to smooth the weighting gain magnitudes [3]. We next experimentally verify the advantage of the proposed approach.

4. EXPERIMENTS AND RESULTS

To test the proposed method experimentally, we first conduct experiments for the purpose of only dereverberation. Secondly, we use the proposed dereverberation scheme as a preprocess to a speech separation algorithm for a reverberant context that is based on the models of the ILD and ITD. We believe this combination would be useful in two ways: firstly, it would be verified that the binaural dereverberation scheme has preserved the ILD and ITD cues which will be utilized in the proceeding speech separation stage; secondly, such a dereverberation scheme could be effectively incorporated in source separation algorithms in highly reverberant scenarios to enhance the overall end-to-end gain of the system. The anechoic speech utterances in all experiments come from the TIMIT database [8]. Real binaural RIRs(BRIRs) from the Aachen impulse response (AIR) database [9] were used in the dereverberation-only experiments while BRIRs measured in real rooms at the University of Surrey [10] were used in the joint dereverberation and source separation experiments. The frame length used was 512 and the frame overlap was 0.75. The other parameter values were the same as in the original works [1, 2, 3].

4.1. Dereverberation-only

Speech files from TIMIT were chosen randomly containing both male and female speakers. In the AIR database, the first set of BRIRs used here were measured in an office room with source-to-microphone distance of 1m and 3m with an RT60 of 0.37s and 0.48s respectively. The BRIRs in the second set were measured in a lecture room with source-to-microphone distance of 2.25m, 5.56m and 10.2m with an RT60 of 0.70s, 0.79s, and 0.83s respectively. Both the LB-RIR and WW-SMOOTH schemes were applied to the observed reverberant signals without any inverse filtering. To give a better insight into the performance of the proposed method we used three objective measures including the signal-to-noise ratio (SNR), segmental SNR (segSNR), and the perceptual evaluation of speech quality (PESQ) [11].

Table 1 summarizes the experimental results in the context of dereverberation. The degraded reverberant speech is referred to as Reverb; LB-RIR in the table means that the signal is enhanced using the LB-RIR method and the bilateral gains are combined using our proposed scheme. Results of the proposed multistage cascaded method follow next. Each value in the table is an average of three randomly selected speech signals from the TIMIT database. We can see that the proposed approach provides an improvement in all the three

Table 1. Mean values of SNR (dB), segmental SNR (segSNR) (dB) and PESQ for three random signals from TIMIT convolved with BRIRs from the Aachen database. RT60s of 0.37, 0.48, 0.70, 0.79, and 0.83 seconds under consideration.

RT60	SNR (dB)			segSNR (dB)			PESQ		
seconds(s)	Reverb	LB-RIR	Proposed	Reverb	LB-RIR	Proposed	Reverb	LB-RIR	Proposed
0.37s	-2.47	-2.07	-1.82	-2.87	-2.23	-1.89	2.78	2.94	3.10
0.48s	-3.26	-2.40	-1.89	-4.89	-3.68	-2.83	2.02	2.11	2.22
0.70s	-3.38	-2.51	-2.11	-4.46	-2.73	-2.13	2.48	2.72	2.75
0.79s	-2.89	-2.14	-1.67	-4.29	-3.15	-2.42	2.05	2.21	2.35
0.83s	-3.55	-2.43	-1.84	-4.93	-3.43	-2.59	2.02	2.24	2.34

evaluation metrics. Over all the RT60s, the proposed method gives a mean SNR gain of 1.13 dB, mean segSNR gain of 1.92 dB, and PESQ improvement of 0.28, compared to LB-RIR which gives an SNR gain of 0.8 dB, segSNR gain 1.24 dB, and a PESQ improvement of 0.17.

4.2. Dereverberation and Source Separation

In this experiment we used the proposed dereverberation scheme as a pre-processing stage to a source separation algorithm utilizing the ILD and IPD cues. We used our recently proposed source separation (SS) algorithm [12] for this purpose. BRIRs used in this experiment [10] were measured in four different rooms with RT60s of 0.32, 0.47, 0.68, and 0.89 seconds. Experiments were also conducted for the LB-RIR method and WW-SMOOTH methods both using our new gain combination scheme, and are referred to as LB-RIR+SS and WW-SMOOTH+SS respectively. The performance of the different methods is measured in terms of PESQ.

The results for this experiment are shown in Fig. 3 where the PESQ values of the different methods are given as a function of RT60. Here SS-only indicates the output of the source separation algorithm without any dereverberation. As is clear from the results, the proposed scheme when used at the first stage before source separation along with suppressing the late reverberation also preserves the binaural cues to enhance the overall gain without deteriorating the speech quality. The Proposed+SS method, over all RT60s, provides average improvement of 0.25, in terms of PESQ, compared to the SS-only. The Proposed+SS method is also consistently better than the LB-RIR and WW-SMOOTH methods at all RT60s.

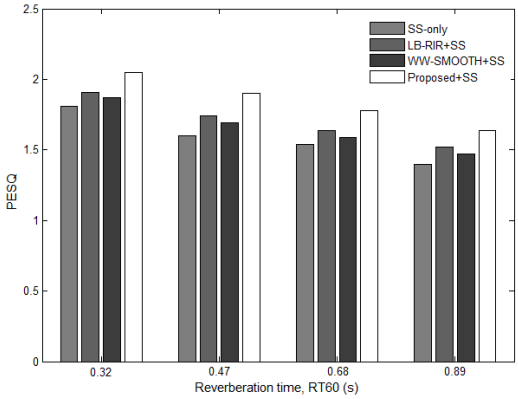

Fig. 3. Results in terms of PESQ for the different methods with varying RT60s. The Proposed+SS scheme performs consistently better at all the RT60s under consideration.

5. RELATION TO PRIOR WORK AND CONCLUSION

The work presented in this paper detailed an efficient binaural speech dereverberation scheme. The proposed scheme is a cascade of three stages utilizing the monaural dereverberation algorithms: [1] and [2] at the first and second stages. In the first two stages we extend the monaural algorithms to the binaural context using a new gain derivation method. The authors in [3] extend the monaural algorithm in [2] to the binaural context by using a delay-and-sum beamformer and form a reference signal. The work in [4] takes a different approach and separately processes the left and right channel signals to produce two gains. Three different gain adaptation schemes are then proposed by taking the minimum, maximum or average of the gains from both the channels. Motivated by the works in [5] and [6], we propose that an overall increased signal enhancement can be achieved with the cascade of nonlinear dereverberation processors rather than using these nonlinear processors individually as in [3] and [4].

This study presented a novel cascaded approach for binaural speech dereverberation. With the proposed cascade and the new gain derivation scheme, increased overall reverberation removal is possible without deteriorating the speech quality. The method provides useful gain improvement in the context of both dereverberation-only and when used as a preprocessing stage to a source separation system applied in a highly reverberant setting. Results in terms of SNR, segSNR and PESQ highlight that the proposed algorithm provides enhancement without deteriorating the speech quality.

The work provides a useful insight into the use of a cascade of binaural dereverberation operations along with the new binaural gain derivation technique. Although this paper utilized three stages in the cascade, the use of more stages or different gain estimation schemes within each stage for an increased signal enhancement warrants further exploration.

6. REFERENCES

[1] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774 – 784, 2006.

[2] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359 – 366, 2001.

[3] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732 – 1745, 2010.

[4] A. Tsilfidis, E. Georganti, and J. Mourjopoulos, "Binaural extension and performance of single-channel spectral subtraction dereverberation algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 1737 – 1740.

[5] D. Arnstein, T. Czerner, and J. Buzzelli, "RFI suppression by cascading nonlinear devices. COMSAT Corp., 1996, US patent no. 5564095, Available: `http://www.google.com/patents/US5564095`," .

[6] D. Arnstein, T. Czerner, and J. Buzzelli, "Broadband signal processing for AJ and RFI reduction in spread spectrum systems," in *Military Communications Conference, 1994. MILCOM '94. Conference Record, 1994 IEEE*, pp. 421 – 429.

[7] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4409 –4412.

[8] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM, 1993. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1W.html," .

[9] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th International Conference on Digital Signal Processing*, 2009, pp. 1 – 5.

[10] C. Hummersone, "Binaural room impulse responses (BRIRs)," *University of Surrey, UK, `http://www.surrey.ac.uk/msr/people/chris_hummersone/BRIRs`.*

[11] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229 – 238, 2008.

[12] M. S. Khan, S. M. Naqvi, A.-Rehman, W. Wang, and J. A. Chambers, "Video-aided model-based source separation in real reverberant rooms," *submitted to IEEE Transactions on Audio, Speech, and Language Processing*, 2012.