Audio Engineering Society

Convention Paper

Presented at the 130th Convention
2011 May 13–16 London, UK

Joint noise and reverberation suppression for speech applications

Elias K. Kokkinis[1], Alexandros Tsilfidis[1], Eleftheria Georganti[1], and John Mourjopoulos[1]

[1] *Audio and Acoustic Technology Group, Department of Electrical and Computer Engineering, University of Patras, 26504, Patras, Greece*

Correspondence should be addressed to Elias K. Kokkinis (ekokkinis@upatras.gr)

ABSTRACT

An algorithm for joint suppression of noise and reverberation from speech signals is presented. The method requires a handclap recording that precedes speech activity. A running kurtosis technique is applied in order to extract an estimation of the late reflections of the room impulse response from the clap while a moving average filter is employed for the noise estimation. Moreover, the excitation signal derived from the Linear Prediction (LP) analysis of the noisy speech along with the estimated power spectrum of the late reflections are used to suppress late reverberation through spectral subtraction while a Wiener filter compensates for the ambient noise. A gain magnitude regularization step is also implemented to reduce overestimation errors. Objective and subjective results show that the proposed method achieves significant speech enhancement in all tested cases.

1. INTRODUCTION

In most practical applications, speech quality and intelligibility are degraded under the effect of room reverberation and ambient noise, while the performance of Automatic Speech Recognition (ASR) systems is significantly compromised. Although the placement of a microphone in close proximity to the speaker may minimize the impact of the above distortions [1], this solution may give rise to other problems such as extreme coloration (*proximity effect*) or may not be desired/feasible in practice for several emerging applications (i.e. immersive communications [2]).

To address this problem, a number of different methods for the suppression of noise ("denoising") and reverberation ("dereverberation") have been proposed. A common method for the suppression of noise is spectral subtraction (SS) [3, 4], where an

estimate of the noise short-time spectrum is subtracted from the microphone's short-time spectrum in order to provide an enhanced signal at the output. Moreover the Wiener filter has been also used for denoising purposes, providing an estimate of the clean signal in the minimum mean-squared error sense. In all cases, the challenge is to accurately estimate the noise short-time spectrum and for a review of relevant techniques the interested reader can refer to [5].

More recently, spectral subtraction has been also employed for the suppression of late reverberation. The room impulse response (RIR) consists of an early and a late reverberant part producing different types of degradation in the speech signal. Hence, most dereverberation techniques compensate separately for the coloration generated from the early reflections and the late reverberant decaying tails. Different SS-based methods that blindly estimate the late reverberation short-time spectrum based either on RIR or signal modelling have been proposed [6, 7, 8, 9, 10]. Other techniques draw on the properties of the reverberant LP residual and the harmonic speech structure [11] or employ auditory modelling [12] in order to achieve reverberation suppression with minimal perceived artifacts.

Given that SS and Wiener filtering are employed for both noise and reverberation suppression it is necessary to examine a combined approach for such tasks. To this end, a method for joint suppression of noise and late reverberation is presented in this work. A recorded handclap along with the excitation signal derived from the Linear Prediction (LP) analysis of the reverberant speech provide the noise and late reverberation spectrum estimations [13]. Then spectral subtraction is applied in order to suppress late reverberation while the ambient noise is suppressed through Wiener post-filtering. A Gain Magnitude Regularization (GMR) step is also implemented to adjust the suppression rate and compensate for processing artifacts [14]. The performance of the proposed technique is evaluated using measured RIRs and noise and it is shown to be robust with respect to the level of noise, reverberation time and clap recording positions, while preserving the quality of the speech signal.

2. METHOD DESCRIPTION

The proposed method is illustrated by the block diagram of Figure 1 and will be explained in detail below.

Consider a microphone placed at a *fixed* position ρ_m and a speaker located at ρ_i, inside a reverberant room, where an ambient noise source (such as a PC fan or an A/C unit) is also present. The speech signal captured by the microphone can be expressed as

$$y_i(k) = s(k) * h_i(k) + n(k) \tag{1}$$

where k is the discrete time index, $h_i(k)$ is the RIR between the speaker and the microphone, $n(k)$ is the signal produced by the noise source and $s(k)$ is the anechoic speech signal. It is assumed here that the noise and speech signals are uncorrelated.

Using LP analysis the speech signal can be modelled as the convolution of an excitation signal $u(k)$ and a speech production filter $h_S(k)$. Furthermore, it is known that a RIR can be decomposed into an early reflection $h_{i,E}(k)$ and a late reverberation $h_{i,L}(k)$ part. Typically the speech production filter is shorter than the early reflection part and hence Eq. 1 can be written as

$$y_i(k) = \sum_{m=0}^{L_B} \sum_{l=0}^{L_S} h_{i,E}(m-l) h_S(l) u(k-m) + \sum_{m=L_B+1}^{L_R} h_{i,L}(m) u(k-m) + n(k) \tag{2}$$

where L_S is the length of $h_S(k)$, L_B is the length of the early reflection part $h_{i,E}(k)$ (or equivalently the *mixing time* t_{mix} expressed in samples) and L_R is the total length of the RIR. Defining the direct signal $d(k)$ and the late reverberant signal $r(k)$, Eq. 2 can be written as

$$y_i(k) = d_i(k) + r_i(k) + n(k) \tag{3}$$

In order to obtain an estimation of the direct (or *clean*) signal, the late reverberant and noise signal or their power spectrums must be estimated. In the proposed method, late reverberation and noise are assumed to be stationary processes and their approximations are provided by a single handclap recording, without the need for a RIR measurement or additional processing of the speech signal (e.g. voice activity detection).

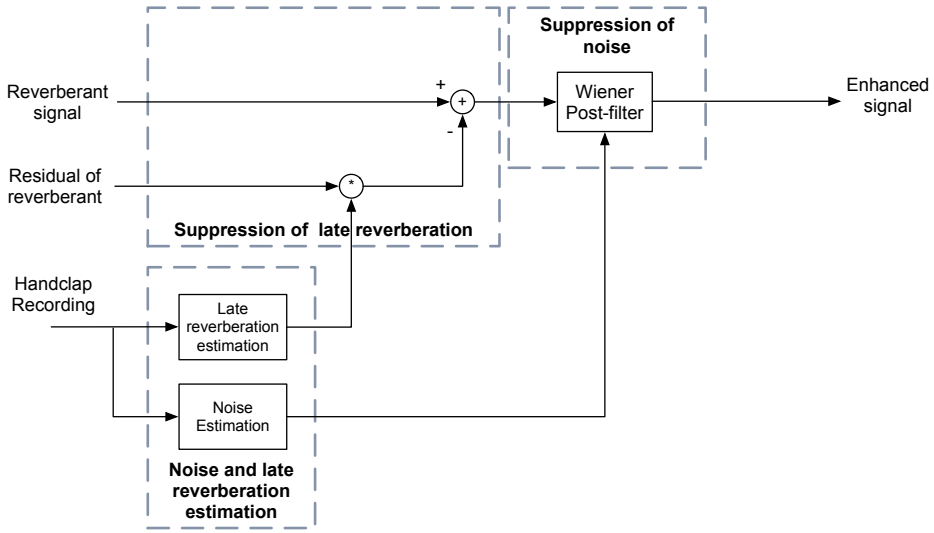

Fig. 1: An illustrative block diagram of the proposed method.

2.1. Estimation of noise and late reverberation from a recorded handclap

Since the energy of late reverberation is considered to be statistically equal in all regions of the room [15], it can be assumed that the power spectral density (PSD) of the late part of a RIR measured for a source at position ρ_i is approximately equal to the PSD of a late part corresponding to a RIR for a source at any other position ρ_j [16]. While a single RIR measurement can lead to fairly good approximation of the late reverberation power spectrum of the room [16], a RIR measurement cannot always be available and a practical approach for the late reverberation estimation is desired.

Consider a handclap recording at position ρ_i

$$c_i(k) = \hat{\delta}(k) * h_i(k) + n(k) \tag{4}$$

where it is assumed that the excitation signal produced by the handclap can approximate an impulse. It has been shown that a recorded clap may differ from a measured RIR in: (i) the low-frequency range, (ii) the details of its spectrum, presenting some sort of spectral coloration and (iii) the lack of the exact spectral energy details for each measurement [17, 18]. Late reverberation arises by definition in the diffuse field and its spectrum is approximately white [15]. It is reasonable to assume that the same applies for the late room response part due to a handclap. In addition, speech signals do not contain significant energy in the low-frequency range. Hence, the above difference can be considered to be insignificant in the context of late reverberation affecting speech signals and the PSD of late reverberation can be efficiently approximated by the PSD of the late part of an in-room handclap recording. Hence

$$c_{i,L}(k) = \hat{h}_{i,L}(k) + n(k) \quad \text{for } k > L_B \tag{5}$$

The early-late RIR boundary L_B or equivalently the "mixing time" t_{mix}, denotes the start of the diffuse field in a room response [15]. It can be usually described as a fixed time interval regardless of the room properties (usually 80 ms), or calculated based on physical quantities such as the room volume [15, 19]. However, the precise evaluation of this early/late reflections boundary from a RIR measurement is a challenging and open research issue [20, 21]. In [16] the authors used a normalized kurtosis approach [15, 20] to calculate the mixing time. The same approach is also used here and applied to the recorded handclap. However, due to the presence of noise and the approximate nature of the obtained "response" the method may digress. Hence if the mixing time calculated by this method is outside a reasonable range (e.g. $50\ ms \leq t_{mix} \leq 500\ ms$ [21]) the static boundary of 80 ms is applied [13].

In order to reduce the effect of noise in the estimated late reverberation, a moving average (MA) with a span of L_M samples is applied to the recorded clap

$$\tilde{c}_{i,L}(k) = \frac{1}{L_M}\left(\sum_{p=1}^{L_M}\hat{h}_{i,L}(k-p) + \sum_{p=1}^{L_M} n(k-p)\right) \quad (6)$$

for $k > L_B$. If the noise source is sufficiently modelled by white Gaussian noise (WGN), then the sum $\sum_{p=1}^{L_M} n(k-p)$ should tend to zero for a large enough L_M. Hence, the PSD of the MA filtered recorded handclap at position ρ_i can be used as an estimation of the late reverberation imprinted on the speech signal.

Since the late part of a RIR is considered to decay exponentially [22], after a certain point L_N noise will dominate the handclap recording

$$c_i(k) \approx n(k) \quad \text{for } k > L_N \quad (7)$$

Eq. 7 can provide an estimation of the noise PSD. L_N can be chosen as the last samples of the handclap recording, equal to the length of the processing frame.

2.2. Suppression of late reverberation

The late reverberant signal can be approximated by using the LP residual $u(k)$ of the captured speech signal $y_i(k)$ and Eq. 6, as

$$\hat{r}_i(k) = u(k) * \tilde{c}_{i,L}(k) \quad (8)$$

By applying a short time Fourier transform (STFT) on Eq. 8 the PSD of the late reverberant signal is obtained and following a spectral subtraction principle, an estimation of the noisy direct signal can be calculated as

$$P_{\tilde{d}_j}(\kappa,\omega) = P_{y_j}(\kappa,\omega) - P_{\hat{r}}(\kappa,\omega) \quad (9)$$

where κ is the STFT frame index and ω the discrete frequency bin index. Note, that since the residual signal does not depend on room properties or source position and the energy of the late reverberation is equal in all room positions, the PSD of the reverberant signal $P_{\hat{r}}(\kappa,\omega)$ can be used to provide estimations of the noisy direct signal in any room position ρ_j.

The above equation can be alternatively expressed as a spectral gain multiplication

$$P_{\tilde{d}_i}(\kappa,\omega) = G_R(\kappa,\omega)P_{y_j}(\kappa,\omega) \quad (10)$$

where

$$G_R(\kappa,\omega) = \frac{P_{y_j}(\kappa,\omega) - P_{\hat{r}}(\kappa,\omega)}{P_{y_j}(\kappa,\omega)} \quad (11)$$

2.3. Suppression of noise

Eq. 7 provides an estimation of the noise signal and with the use of STFT a sufficient approximation of the noise power spectrum can be derived. Hence an approximate Wiener filter can be calculated in the time-frequency domain to facilitate the suppression of noise in the estimated direct signal. The Wiener filter can be expressed as

$$G_N(\kappa,\omega) = \frac{P_{\tilde{d}_i}(\kappa,\omega)}{P_{\tilde{d}_i}(\kappa,\omega) + P_{\hat{n}}(\kappa,\omega)} \quad (12)$$

Applying the Wiener filter to the output of the spectral subtraction process described in Eq. 11

$$P_{\hat{d}_i}(\kappa,\omega) = G_N(\kappa,\omega)P_{\tilde{d}_i}(\kappa,\omega) \quad (13)$$

and by substituting Eq. 10 to 13, the PSD of the estimated clean direct signal, i.e. the enhanced speech signal with suppressed noise and late reverberation becomes

$$P_{\hat{d}_i}(\kappa,\omega) = G_J(\kappa,\omega)P_{y_i}(\kappa,\omega) \quad (14)$$

2.4. Gain Magnitude Regularization

In order to compensate for overestimation errors and prevent the generation of musical noise and other artifacts commonly presented in similar methods, the total applied spectral gain is constrained through the following equation:

$$G_J(\kappa,\omega) = \begin{cases} \dfrac{G_J(\kappa,\omega)-\theta}{r} + \theta & \text{when } \zeta < \zeta_{th} \\ & \text{and } G_J(\omega,j) < \theta \\ G_J(\kappa,\omega) & \text{otherwise} \end{cases} \quad (15)$$

where

$$\zeta = \frac{\sum_{\omega=1}^{\Omega} G_J(\kappa,\omega)P_{y_j}(\kappa,\omega)}{\sum_{\omega=1}^{\Omega} P_{y_j}(\kappa,\omega)} \quad (16)$$

where θ is the threshold for applying the gain constraints, r is a regularization ratio, ζ is the power

Room	Type	Volume (m^3)	RT_{60} (sec)
1	Lect. room	555	0.84
2	Conf. hall	1292	1.00

Table 1: Summary of the type and acoustical properties of the rooms where the RIRs were measured.

ratio between the enhanced and the unprocessed signal, ζ_{th} is a threshold for the detection of regions where speech has low power when compared to noise and/or reverberation, and finally Ω is the number of frequency bins.

3. TESTS AND RESULTS

3.1. Measurements

Eight anechoic speech recordings, both male and female (sampled at 44.1 kHz) were convolved with measured RIRs in two different rooms (see Table 1) at various distances (0.5m, 1m, 3m). In addition, handclaps in different (random) positions in those rooms and the noise of a typical office A/C unit were also recorded. From these data, the noisy and reverberated speech signals were generated. Note also that in each case noise has been added to the clap recordings at the same SNR as that of the noisy speech signal. The processing was applied for frame size of 4096 samples with a 50% overlap, $\theta = 0.25$ and $\zeta_{th} = 0.8$, the value of the regularization ratio r was 3 and the LP analysis order was 13.

The performance of the proposed method was assessed by calculating the improvement between the degraded and enhanced signals in terms of Signal to Noise Ratio (SNR) and Perceptual Speech Quality (PESQ) [5]. Note that the SNR takes into account both noise and reverberation and hence it can assess the performance of the combined method, as well as the performance of each part of the method, namely the suppression of noise only or late reverberation only.

3.2. Performance comparison for suppression of noise or late reverberation only and joint suppression

In this section the performance when only noise or reverberation suppression is applied will be examined in comparison to the performance of the joint suppression method. For this, reverberant recordings with the noise source at SNR=15dB were created using the RIRs of room 1, as well as clap recordings with the same SNR. The results are shown in Figures 2 and 3.

In any of the examined cases, the joint suppression approach provides greater improvement in terms of both SNR and PESQ than when only one of the degrading effects are addressed. Furthermore, the improvement is fairly consistent regardless of which clap recording is used to obtain the spectrum estimations. Finally, it is interesting to note that the performance is also quite similar for the case of a speaker at 0.5m and 1.0m, indicating that at typical office distances between the speaker and the microphone the method performs quite well, while performance is a little lower for a speaker at 3.0m especially in terms of PESQ (Fig. 3(c)).

3.3. Performance for variable SNR and reverberation time

Here the performance of the joint suppression method will be examined for rooms with different reverberation times and recordings (both speech and claps) at different SNRs.

The results for room 1 are shown in Figures 4 and 5. The performance improvement decreases in terms of SNR and increases in terms of PESQ with an increasing SNR of the microphone and clap recordings. Additionally, the performance of the proposed method is consistent for a given SNR regardless of the clap recording or the speaker position.

The results for room 2, with a longer reverberation time, are shown in Figures 6 and 7. The improvement in terms of SNR and PESQ follows the same trends as for room 1.

4. CONCLUSIONS

A method for the joint suppression of noise and late reverberation from speech signals was presented, considering practical hands-free applications in real environments in the presence of typical noise sources. The suppression of late reverberation was based on a spectral subtraction technique and the suppression of noise on a Wiener post-filter. The power spectrum estimations of noise and late reverberation were obtained simultaneously from a handclap recording making use of the fact that the energy

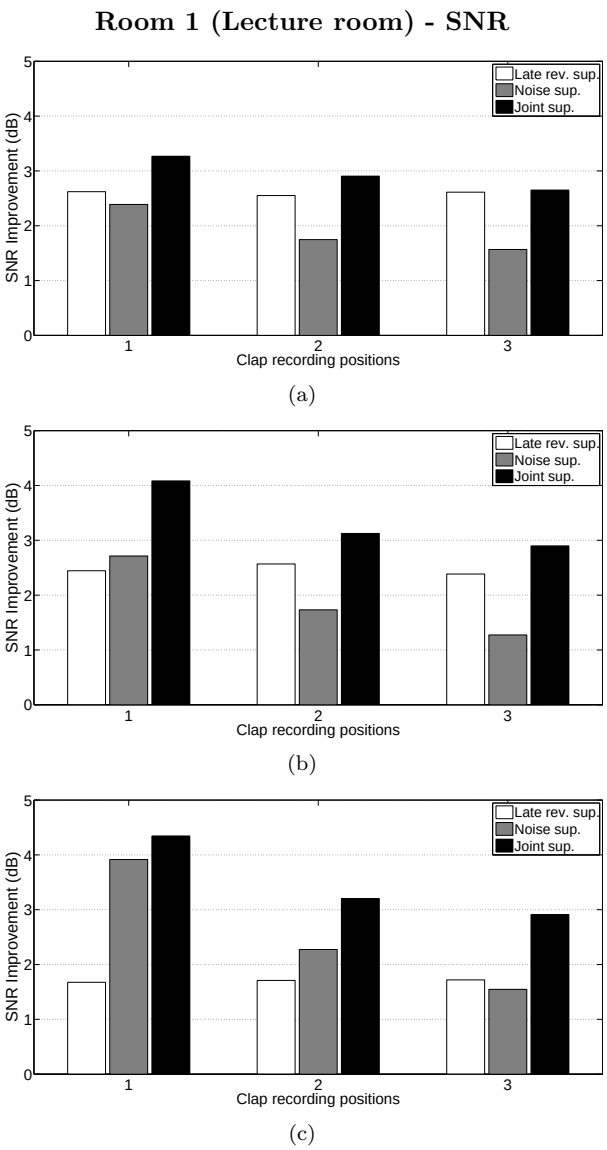

Fig. 2: The average performance of the separate parts and the combined method in terms of SNR for a RIR measured in room 1 at a source-microphone distance of (a) 0.5m, (b) 1.0m and (c) 3.0m. The degraded signals and the clap recordings have a SNR of 15dB.

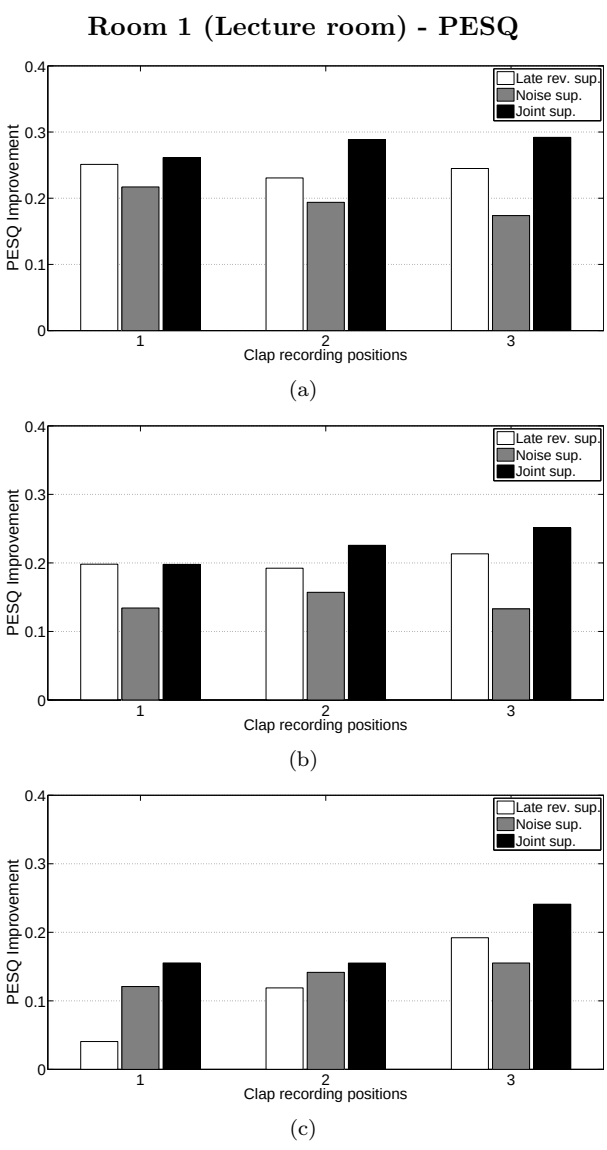

Fig. 3: The average performance of the separate parts and the combined method in terms of PESQ for a RIR measured in room 1 at a source-microphone distance of (a) 0.5m, (b) 1.0m and (c) 3.0m. The degraded signals and the clap recordings have a SNR of 15dB.

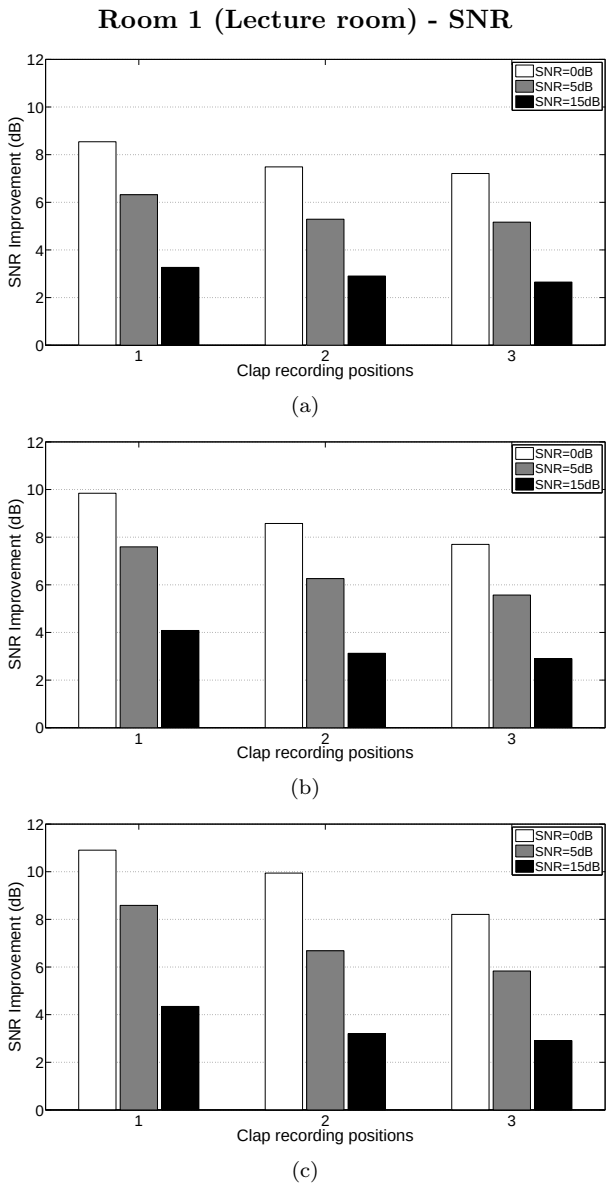

Fig. 4: The average performance of the joint suppression method in terms of SNR for a RIR measured in room 1 at a source-microphone distance of (a) 0.5m, (b) 1m and (c) 3m.

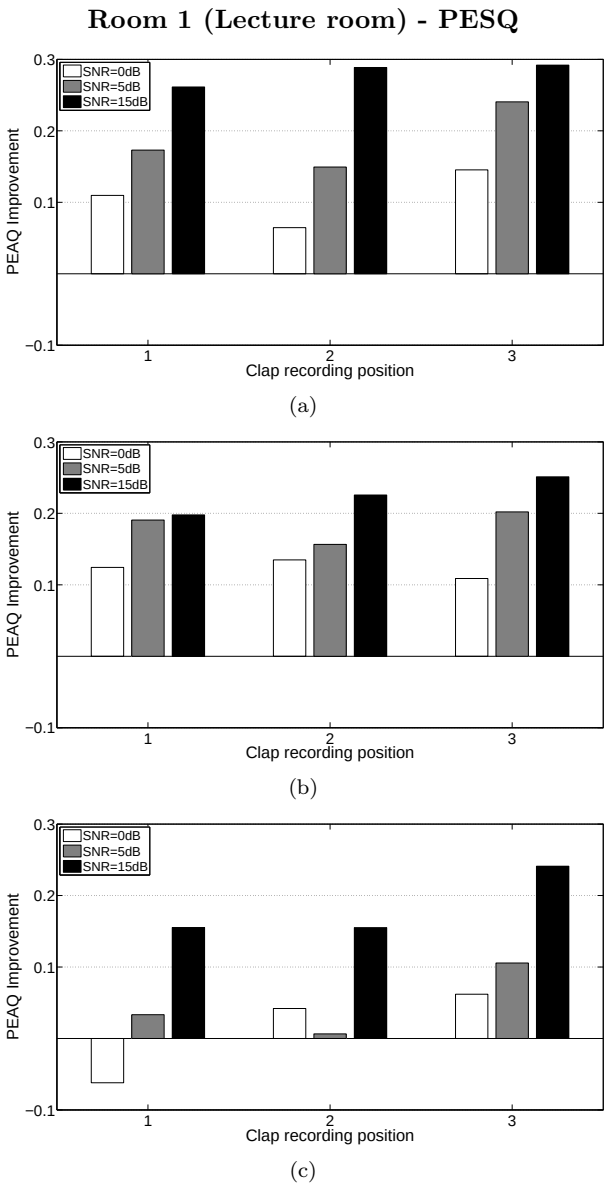

Fig. 5: The average performance of the joint suppression method in terms of PESQ for a RIR measured in room 1 at a source-microphone distance of (a) 0.5m, (b) 1m and (c) 3m.

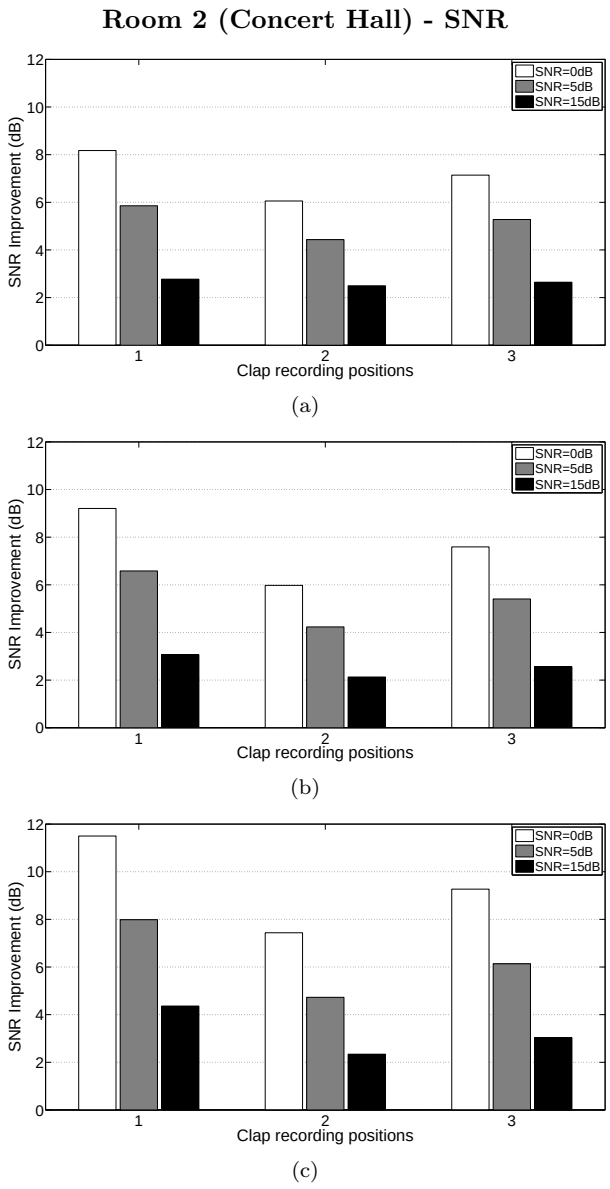

Fig. 6: The average performance of the joint suppression method in terms of SNR for a RIR measured in room 2 at a source-microphone distance of (a) 0.5m, (b) 1m and (c) 3m.

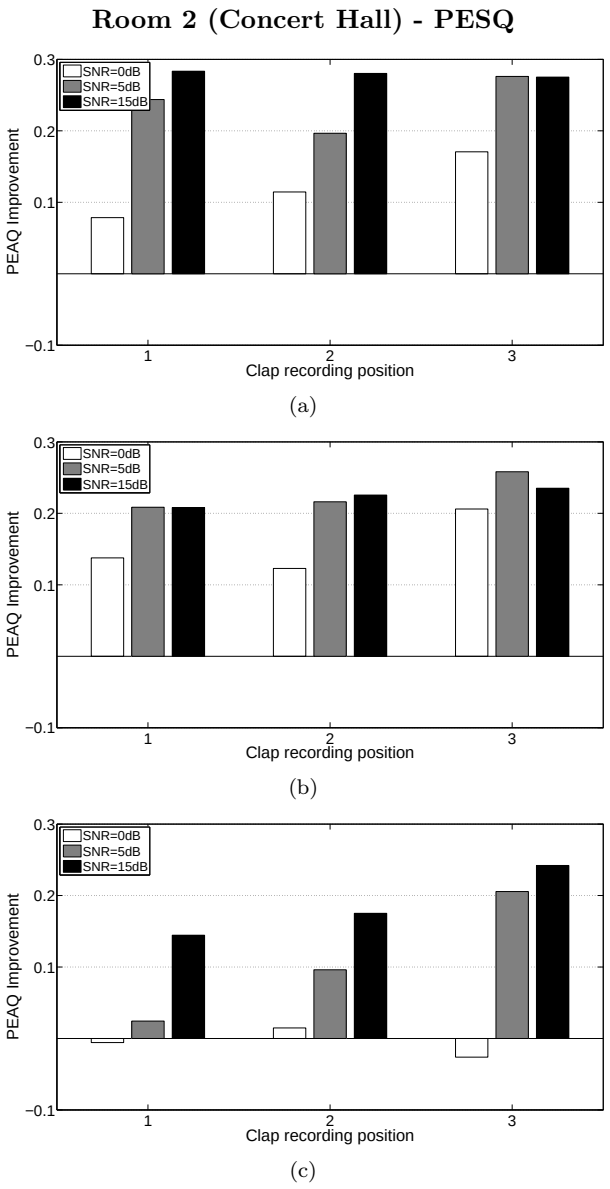

Fig. 7: The average performance of the joint suppression method in terms of PESQ for a RIR measured in room 2 at a source-microphone distance of (a) 0.5m, (b) 1m and (c) 3m.

of the late reverberation is approximately equal in all room positions and the exponential decay of RIRs.

The results presented here indicate that the method does not introduce significant distortion and preserves speech quality, while successfully suppressing late reverberation and noise. The clap recordings have been proved to be a simple and efficient way to obtain simultaneous estimations of noise and late reverberation, without any requirement for special signals and equipment. The method was also shown to be robust with regard to reverberation time, speaker position and handclap recording position. The proposed method is suitable for practical applications where real hands-free acquisition of speech is required providing a simple and efficient enhancement framework.

ACKNOWLEDGEMENT

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 (ESPA) 2007-2013, according to Contract no. MICRO2-07/E-II-A.

5. REFERENCES

[1] E. K. Kokkinis and J. Mourjopoulos. Identification of a room impulse response using a close-microphone reference signal. In *128th AES Convention 128*, May 2010.

[2] Yiteng Huang, Jingdong Chen, and J. Benesty. Immersive audio schemes. *IEEE Signal Processing Magazine*, 28(1):20–32, January 2011.

[3] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Process.*, 27(2):113 – 120, April 1979.

[4] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1979.

[5] P Loizou. *Speech enhancement: theory and practice*. CRC Press, 1st edition, 2007.

[6] K. Lebart and J. Boucher. A new method based on spectral subtraction for speech dereverberation. *Acta Acust. Acust.*, 87:359–366, 2001.

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

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

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

[10] R. Gomez and T Kawahara. Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood. *IEEE Trans. Audio, Speech and Lang. Process.*, 18:1708–1716, 2010.

[11] B. Yegnanarayana and P.S. Murthy. Enhancement of reverberant speech using lp residual signal. *IEEE Transactions on Speech and Audio Processing*, 8(3):267 –281, May 2000.

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

[13] A Tsilfidis, E Georganti, E Kokkinis, and J Mourjopoulos. Speech dereverberation based on a recorded handclap. In *Digital Signal Processing Conf. (DSP) (submitted)*, Corfu, Greece, 2011.

[14] A Tsilfidis, E Georganti, and J Mourjopoulos. Binaural extension and performance of single-channel spectral subtraction dereverberation algorithms. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

[15] B. Blesser. An interdisciplinary synthesis of reverberation viewpoints. *J. Aud. Eng. Soc.*, 49(10):867–903, 2001.

[16] A. Tsilfidis, E. K. Kokkinis, and J. Mourjopoulos. Suppression of late reverberation at multiple speaker positions utilizing a single impulse response measurement. In *Forum Acusticum (accepted)*, Aalborg, Denmark, 2011.

[17] Dragana Sumarac-Pavlovic, Miomir Mijic, and Husnija Kurtovic. A simple impulse sound source for measurements in room acoustics. *Applied Acoustics*, 69(4):378 – 383, 2008.

[18] Bruno H. Repp. The sound of two hands clapping: An exploratory study. *J. Acoust. Soc. Amer.*, 81(4):1100–1109, 1987.

[19] G. Defrance and J.-D. Polack. Measuring the mixing time in auditoria. In *Proc. of the Acoustics '08*, pages 3871–3876, Paris, France, June-July 2008.

[20] R. Stewart and M. Sandler. Statistical measures of early reflections of room impulse responses. In *Proc. of the 10th Int. Conf. on Digital Audio Effects (DAFx-07)*, pages 1–4, Bordeaux, France, September 2007.

[21] Takayuki Hidaka, Yoshinari Yamada, and Takehiko Nakagawa. A new definition of boundary point between early reflections and late reverberation in room impulse responses. *J. Acoust. Soc. Amer.*, 122(1):326–332, 2007.

[22] J. D. Polack. Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics. *Applied Acoustics*, (38), 1993.