SPEECH DEREVERBERATION BASED ON A RECORDED HANDCLAP

Alexandros Tsilfidis, Eleftheria Georganti, Elias K. Kokkinis, John Mourjopoulos

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

ABSTRACT

A semi-blind framework for the suppression of late speech reverberation is presented. The method is based on spectral subtraction and utilizes a simple recorded handclap to estimate the Power Spectral Density (PSD) of late reverberation. A statistical analysis of measured Room Impulse Responses (RIRs) and recorded handclaps demonstrates the sufficiency of the above estimation. Dereverberation results show that the proposed technique achieves significant reverberation suppression at multiple speaker positions in each room without compromising the quality of the estimated signals.

***Index Terms*—** dereverberation, moving speaker, handclap, room impulse response, statistics

1. INTRODUCTION

Room reverberation is often considered as an unwanted distortion that reduces speech intelligibility and deteriorates the performance of many applications in communications and in Automatic Speech Recognition (ASR) systems. When the microphone that captures the speech signal is placed close to the speaker the speech degradation is minimized, since in effect the close microphone technique reduces the effect of room acoustics on the captured signal [1]. However, in many real life scenarios and applications, such close-microphone implementations are not feasible and dereverberation post processing is required in order to enhance the captured speech signals. Most current dereverberation methods handle separately the spectral degradation produced by the early reflections and the noise-like late reverberation effect. Late reverberation is to a large extent responsible for the speech intelligibility reduction and ASR performance degradation and consequently several methods have been proposed in order to compensate for this particular type of distortion; many among them employing spectral subtraction (e.g. [2–7]). Some of these techniques are blind, i.e. no reference measurements of room acoustics, typically via the Room Impulse Response (RIR), are required. However, accurate dereverberation requires prior RIR measurements which can be impractical in many real-life situations. Furthermore, given that RIRs vary drastically for different source/receiver positions, individual RIRs measurements must be also undertaken to compensate for all other source/receiver positions in the room.

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".

Recently, Tsilfidis et al. [8] have introduced a late reverberation suppression method based on a single RIR measurement. This method is also based on spectral subtraction and estimates the late reverberation's spectral magnitude utilizing the late part of a measured impulse response and the excitation signal derived from the Linear Prediction (LP) analysis of the reverberant signal. The late RIR part is assumed to be a wide-sense stationary process. The above estimation was found to be valid for multiple speaker positions and the method achieves efficient suppression of late reverberation when moving sources are involved.

However, a RIR measurement is not always feasible since in occupied rooms the recording of an excitation signal may be disturbing for the occupants. In the present work, the above technique is extended in a semi-blind framework and in order to derive the required approximation of the late reflections Power Spectral Density (PSD), a single recorded handclap is utilized. This is a flexible option when a RIR measurement is not feasible since a handclap recording can provide a reasonable RIR approximation [9]. Comparing to a properly measured RIR, the handclaps usually have a more pronounced radiation directivity and contain less energy in the low frequencies [10] whilst presenting some spectral coloration, largely varying between different handclap signals [10, 11]. However, for this dereverberation method, they can provide an acceptable approximation of the late reverberant PSD (see Section 4.2) and moreover, the proposed approach provides additional compensation for estimation errors through a Gain Magnitude Reguralization step [8]. The results show significant late reverberation suppression in various Reverberation Time (RT) conditions and source/receiver distances, indicating that the proposed approach is appropriate for real life applications.

2. METHOD DESCRIPTION

In room acoustics, the RIR $h(n)$ can be decomposed to an early and a late reverberation part:

$$h(n) = h_{early}(n) + h_{late}(n) \tag{1}$$

where n denotes the discrete time index. These two RIR parts introduce different effects on the reverberant speech signal $y(n)$, i.e.:

$$y(n) = \sum_{m=0}^{L_b} h_{early}(m)s(n-m) + \sum_{m=L_b}^{L_r} h_{late}(m)s(n-m) \tag{2}$$

978-1-4577-0274-7/11/$26.00 ©2011 IEEE

where L_b is the early-late reflections boundary, L_r is the length of the RIR and $s(n)$ the anechoic signal. Using the LP analysis, a speech signal is modelled as the convolution of an excitation signal $u(n)$ and a speech production filter $h_s(n)$. Hence, Eq. 2 can be written as:

$$y(n) = \sum_{m=0}^{L_b} \sum_{l=0}^{L_s} h_{early}(m-l)h_s(l)u(n-m) + \sum_{m=L_b+1}^{L_r} \sum_{l=0}^{L_s} h_{late}(m-l)h_s(l)u(n-m) \quad (3)$$

Assuming that the length of the speech production filter (L_s) is shorter than the length of $h_{early}(n)$ (i.e. $L_s < L_b$) [12], Eq. 3 becomes:

$$y(n) = \sum_{m=0}^{L_b} \sum_{l=0}^{L_s} h_{early}(m-l)h_s(l)u(n-m) + \sum_{m=L_b+1}^{L_r} h_{late}(m)u(n-m) \quad (4)$$

The energy of the late reverberation is statistically equal in all regions of the room [13]. Hence, in [8] it is assumed that the PSD of the late part of a single RIR from a certain room position ρ_0 can be used as an approximation of the PSD of the late part of the RIR for any other room position ρ_i:

$$|H^i_{late}(\kappa,\omega)|^2 \approx |H^0_{late}(\kappa,\omega)|^2 \quad \forall i \quad (5)$$

where κ and ω are the time frame and the frequency bin indices respectively and $H^0_{late}(\kappa,\omega)$ and $H^i_{late}(\kappa,\omega)$ are the Short Time Fourier Transforms (STFT) of the late part of the RIRs in positions ρ_0 and ρ_i respectively. Here, it is further assumed that the PSD of the late reverberation in position ρ_0 can be approximated from the PSD of the late part of a handclap recording $|C^0_{late}(\kappa,\omega)|^2$ in the same position, i.e.:

$$|H^0_{late}(\kappa,\omega)|^2 \approx |C^0_{late}(\kappa,\omega)|^2 \quad (6)$$

The validity of the above approximation is further discussed in Sections 3 and 4.2. Following the spectral subtraction principle and based on Eq. 4, 5 and 6, an estimation of the direct signal's power spectrum can be derived:

$$|\hat{S}^i(\kappa,\omega)|^2 = |Y^i(\kappa,\omega)|^2 - |C^0_{late}(\kappa,\omega)|^2|U^i(\kappa,\omega)|^2 \quad (7)$$

where $|\hat{S}^i(\kappa,\omega)|^2$ and $|Y^i(\kappa,\omega)|^2$ are the PSD of the estimated clean and the reverberant signals in position ρ_i respectively and $U^i(\kappa,\omega)$ is the STFT of the LP residual of the reverberant signal. The subtraction in Eq. 7 can be alternatively formulated as a spectral gain multiplication:

$$|\hat{S}^i(\kappa,\omega)|^2 = G(\kappa,\omega)|Y^i(\kappa,\omega)|^2 \quad (8)$$

where

$$G(\kappa,\omega) = \frac{|Y^i(\kappa,\omega)|^2 - |C^0_{late}(\kappa,\omega)|^2|U^i(\kappa,\omega)|^2}{|Y^i(\kappa,\omega)|^2} \quad (9)$$

In order to compensate for overestimation errors the above gain can be constrained through a Gain Magnitude Regularization (GMR) technique [8] ; the gain of Eq. 9 being constrained as:

$$G(\kappa,\omega) = \begin{cases} \frac{G(\kappa,\omega)-\theta}{r} + \theta & \text{when } \zeta < \zeta_{th} \\ G(\kappa,\omega) & \text{otherwise} \end{cases} \quad (10)$$

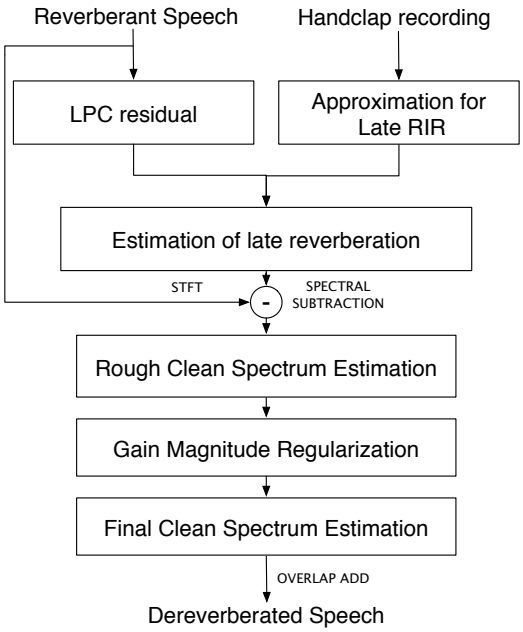


Fig. 1. Signal flow of the proposed method

$$\zeta = \frac{\sum_{\omega=1}^{\Omega} G(\kappa,\omega)|Y^i(\kappa,\omega)|^2}{\sum_{\omega=1}^{\Omega} |Y^i(\kappa,\omega)|^2} \quad (11)$$

where θ is the threshold for applying the gain constraints, r is a regularization ratio, ζ is the power ratio between the enhanced and the reference signal, ζ_{th} a low SRR detector threshold and Ω is the frame size. The signal flow of the proposed dereverberation approach is shown in Fig. 1.

3. ESTIMATING THE LATE REFLECTIONS' POWER SPECTRAL DENSITY (PSD) FROM A HANDCLAP RECORDING

Previous work has 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 [10, 11]. Late reverberation arises by definition in the diffuse field and its spectrum is approximately white [13]. 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 reflections can be efficiently approximated by the PSD of the late part of an in-room handclap recording. In order to experimentally verify the above assumptions, a statistical measure that compares the properties of the PSD of handclaps and RIRs is introduced in Section 3.2; this approach being based on previous studies of the authors related to the statistical analysis of RIRs and signals [14, 15].

3.1. Early-late reverberation boundary

The early-late RIR reverberation boundary, namely the "mixing time" t_{mix}, denotes the start of the diffuse field in a room response [13]. 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 [13, 16]. However, the precise evaluation of this early/late reflections boundary from a RIR measurement is a challenging and open research issue [17, 18]. In [8] the authors derive the mixing time through a normalized kurtosis approach [13, 17]. The measured RIR $h^0(n)$ is partitioned in non-overlapping frames of length L_k and for the kth frame $\boldsymbol{h}_k^0$ the normalized kurtosis is calculated as follows:

$$Kurt[\boldsymbol{h}_k^0] = \frac{E[\boldsymbol{h}_k^0 - \mu]^4}{\sigma^4} - 3 \tag{12}$$

The mixing time in samples is defined as $L_b = k_{min} L_\kappa$ and k_{min} is given by

$$k_{min} = \arg\min\{Kurt[\boldsymbol{h}_k^0]\} \tag{13}$$

The above method is also used here for the determination of the late part of an in-room handclap recording, but due to the noisier nature of such handclap recordings it has been found that it sometimes fails to provide a robust mixing time estimation. Hence, the normalized kurtosis technique is initially employed and if the derived t_{mix} value is within a reasonable range (e.g. $50\ ms \leq t_{mix} \leq 500\ ms$ [18]), then the derived value is used; in any other case the static threshold of 80 ms is applied.

3.2. Statistical measure

Denoting a room transfer function (RTF) derived from a RIR measured at position ρ_i as $H^i(\omega)$, the statistics of the PSD can be calculated either for the full frequency range or for specific sub-bands, i.e. using fractional-octave bandwidths, typically derived from a bank of cascaded $1/f_r$-octave-band filters that divide octaves into f_r subbands. Denoting by $\omega_{l,\phi}$ and $\omega_{u,\phi}$ the lower and upper band edge frequencies of the frequency subband ϕ, then the standard deviation of the PSD of the RIR for subband ϕ, is defined as:

$$\sigma^i_{\phi,h} = \left[\frac{1}{\omega_{u,\phi} - \omega_{l,\phi} + 1} \sum_{\omega=\omega_{l,\phi}}^{\omega_{u,\phi}} (|H^i(\omega)|^2 - \mu^i_{\phi,h})^2 \right]^{\frac{1}{2}} \tag{14}$$

where $\mu^i_{\phi,h}$ is the PSD mean of the RIR for the specific frequency band at position ρ_i, given by:

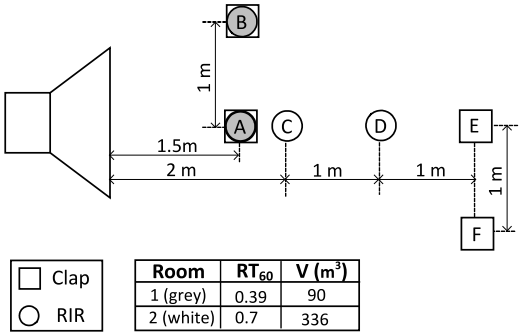
$$\mu^i_{\phi,h} = \frac{1}{\omega_{u,\phi} - \omega_{l,\phi} + 1} \sum_{\omega=\omega_{l,\phi}}^{\omega_{u,\phi}} |H^i(\omega)|^2 \tag{15}$$

The same statistical measure can be calculated for an in-room recorded handclap:

$$\sigma^i_{\phi,c} = \left[\frac{1}{\omega_{u,\phi} - \omega_{l,\phi} + 1} \sum_{\omega=\omega_{l,\phi}}^{\omega_{u,\phi}} (|C^i(\omega)|^2 - \mu^i_{\phi,c})^2 \right]^{\frac{1}{2}} \tag{16}$$

where $C^i(\omega)$ is the spectrum of the recorded handclap and $\mu^i_{\phi,c}$ is the PSD mean of the handclap for the specific frequency band at position ρ_i, given by:

$$\mu^i_{\phi,c} = \frac{1}{\omega_{u,\phi} - \omega_{l,\phi} + 1} \sum_{\omega=\omega_{l,\phi}}^{\omega_{u,\phi}} |C^i(\omega)|^2 \tag{17}$$



Room	RT_{60}	V (m^3)
1 (grey)	0.39	90
2 (white)	0.7	336

Fig. 2. Illustration of the measurement setup for the RIR and handclaps.

In a similar manner, the PSD standard deviation can be calculated only for the late part of a RIR ($\sigma^i_{\phi,h_{late}}$) or a handclap ($\sigma^i_{\phi,c_{late}}$) for the frequency subband ϕ.

4. TESTS AND RESULTS

4.1. Measurements

Sixteen speech phrases taken from the TIMIT database (sampled at 16 kHz with 16 bit precision) were convolved with real RIRs measured in two different rooms: (a) a listening room (Room 1) and (b) a lecture hall (Room 2). In addition, handclaps in different positions in those rooms were also recorded. The room acoustical properties and the setup of the measurements are presented in Fig. 2. Squares denote the positions where handclaps were recorded whereas circles symbolize the positions where the impulse responses were measured. The authors conducted unofficial listening tests in order to choose the parameters for the proposed method. Hence, the dereverberation processing was applied for frame size of 1024 samples with a 25% overlap (applying also a zero padding of 1024 samples), the thresholds θ and ζ_{th} were set at 0.4, the value of the regularization ratio r was 6 and the LP analysis order was 13.

4.2. Validation of using handclap recordings

4.2.1. Position dependent comparison

In order to verify the assumption that the PSD of late reflections (late RIR part) can be efficiently approximated by the PSD of the late part of an in-room handclap recording, the approach described in Section 3 was followed. Consequently, the statistical discrepancy for the frequency subband ϕ is measured (in dB) using the absolute statistical error, which is given by:

$$\epsilon^i_{\phi,h\_c} = 20 * \log_{10} |\sigma^i_{\phi,h} - \sigma^i_{\phi,c}| \tag{18}$$

where $\sigma_{\phi,h}$ and $\sigma_{\phi,c}$ correspond to the standard deviation values (see Eq. 14 and Eq. 16) of the PSD of the RIR and the in-room handclap recording, respectively. This error was derived for the late part of the handclap and RIR, as:

$$\epsilon^i_{\phi,h\_c,late} = 20 * \log_{10} |\sigma^i_{\phi,h_{late}} - \sigma^i_{\phi,c_{late}}| \tag{19}$$

In Fig. 3, the absolute error of the standard deviation (in dB) between the RIRs and the handclap recordings, ϵ^i_{ϕ,h_c} using 1/3 octave band analysis is shown. In the figure, the error of RIR and handclap transfer functions at position "A" (ρ_A) of Room 1 are shown with

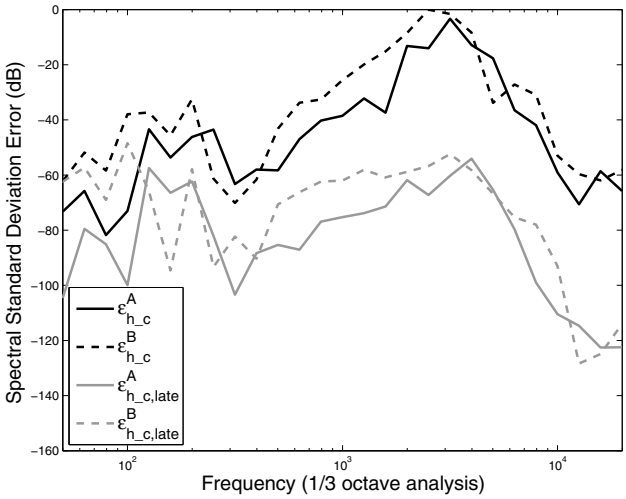


Fig. 3. Standard deviation absolute error between the measured RIRs and in-room handclap recordings for positions "A" and "B" in Room 1 (see Section 4.1) using 1/3 octave band analysis. The error is calculated for the total duration and for the late parts of the RIRs and handclaps.

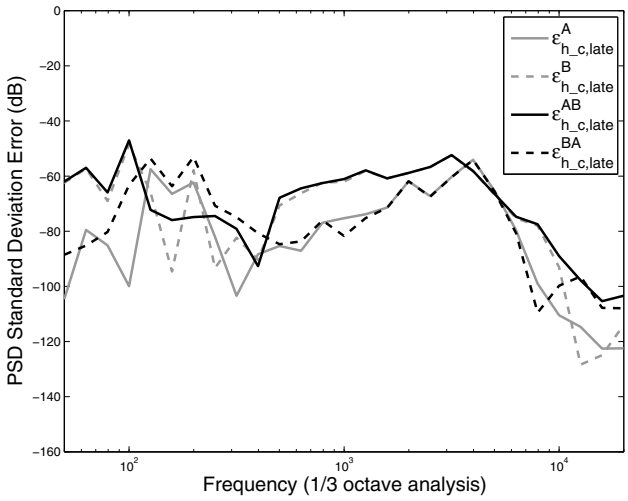


Fig. 4. Standard deviation absolute error between the late part of the RIR measured in positions "A" and "B" and the late parts of each of the handclaps recorded at the same positions (see Section 2) using 1/3 octave band analysis.

black solid line and at position "B" (ρ_B) with black dashed line. It can be noted that that the two curves follow similar trends and their values are quite close, presenting a peak at the region of 2 to 3 kHz. In the same figure, the error between the standard deviation values of the PSD for the late parts of the RIRs and in-room handclap recordings ($\epsilon^{i}_{\phi,h\_c,late}$) is also plotted. It is evident that these errors present much lower values compared to the errors calculated using the total duration of the RIRs and the in-room clap recordings. This is an indication that the statistical properties of the late parts of the RIRs and in-room handclaps present significant statistical similarities.

4.2.2. Position independent comparison

The same analysis, but now focusing at the statistical properties of the late parts of RIRs and handclap recordings, is used for examining variations across different room positions. In Fig. 4 the absolute error of the standard deviation between the late part of the RIR, measured in positions "A" and "B", is compared to the late parts of each of the handclaps recorded at the same positions using 1/3 octave band analysis. It can be seen that the error remains similar when comparing the RIR measured at position "A" with the handclap recorded either at position "A" or "B" (denoted as $\epsilon^{A}_{h\_c,late}$ and $\epsilon^{AB}_{h\_c,late}$ respectively) and this is also valid for the RIR measured at position "B" ($\epsilon^{B}_{h\_c,late}$ and $\epsilon^{BA}_{h\_c,late}$). This indicates that the PSD of the late part of a handclap recorded in a certain room position can be used as an approximation of the PSD of the late part of the RIR for another position.

4.3. Dereverberation Results

Here, the results from the dereverberation processing (see Fig. 1) of the speech signals using the procedure described in Section 2 are given. The mean Signal to Reverberation Ratio (SRR) difference (e.g. [5]) between the reverberant and the estimated clean signals is shown in Fig. 5. Two source-receiver positions were tested in each room. Note that positive SRR differences denote the absolute

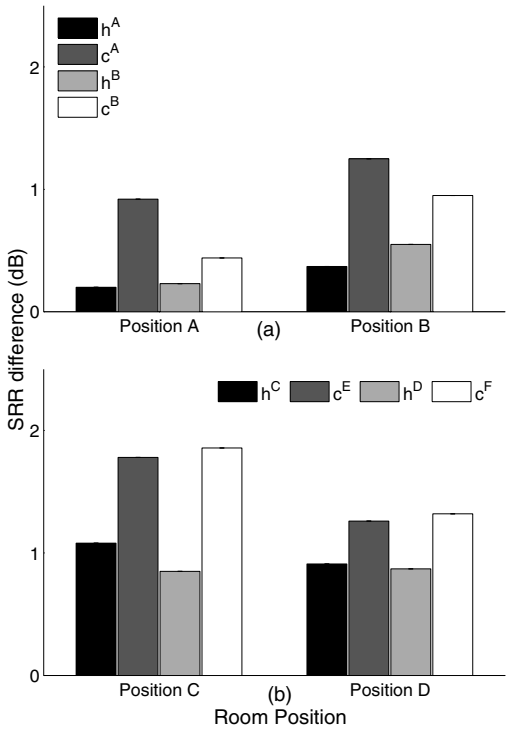


Fig. 5. Mean SRR difference of the reverberant and the estimated clean signals for (a) Room 1, (b) Room 2.

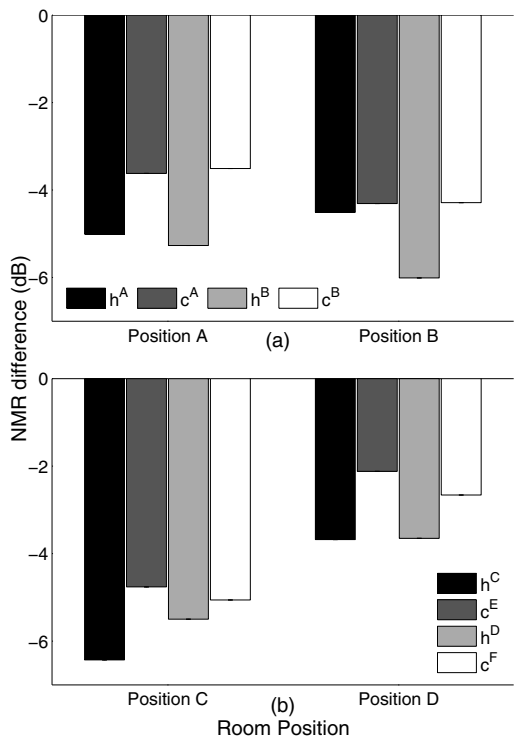


Fig. 6. Mean NMR difference between the estimated clean and the corresponding reverberant signals for (a) Room 1, (b) Room 2.

improvement when compared to the reverberant signals. In Room 1, the proposed method was evaluated using as a reference a RIR and a clap recorded at position "A" (h^A and c^A respectively) and a RIR and a clap recorded at position "B" (h^B and c^B respectively). In Room 2 the method was evaluated utilizing two RIRs measured in positions "C" and "D" (h^C and h^D respectively) and two claps recorded in positions "E" and "F" (c^E and c^F respectively). In all cases an improvement in terms of SRR was observed, this improvement being greater for the larger room (Room 2). In addition, it seems that the use of the actual RIRs as reference, reduces the late reverberation suppression in all positions.

The Noise to Mask Ratio (NMR) measure was also used to assess the quality of the clean signal estimations [3, 4]. The NMR is an objective measure that evaluates the audible (non-masked) noise components and lower NMR values denote better signal quality. The mean NMR difference (e.g. [5]) between the estimated clean and the corresponding reverberant signals for the same experimental conditions as above is shown in Fig. 6. Note that negative NMR difference values denote the relative improvement when compared to the reverberant signals. The recorded claps achieved significant NMR improvement in all tested cases; however slightly better results were obtained when the actual RIRs were used as a reference. This indicates that the greater SRR values achieved from the reference claps (see Fig. 5) may be to a some extent due to an overestimation of the late reverberant PSD. However, the substantial overall improvement both in NMR and SRR shows that the proposed approach achieves sufficient suppression of late reverberation and improves the quality of the produced signals, as was also confirmed after several informal listening tests [1] performed by the authors.

5. CONCLUSION

It has been shown that when a single in-room recorded handclap is used to approximate the power spectral density of late reverberation, then the measured signal provides sufficient features to allow spectral subtraction-based speech dereverberation. A RIR/handclap statistical analysis has been presented to define the boundaries between early/late room response parts. It has been also shown that the comparison between measured RIRs and recorded handclaps validates the above assumptions for the late response properties. Furthermore, by recording a single handclap it appears that sufficient robustness can be achieved with respect to different source/receiver positions within the room. In all cases, the proposed method has shown to achieve sufficient SRR and NMR improvements when employed for late speech reverberation suppression. In future work, the authors will further investigate the validity of the handclap approximation in different rooms providing an additional evaluation of the proposed approach.

6. REFERENCES

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

[2] K Lebart, J Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, pp. 359–366, 2001.

[3] 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.

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

[5] A. Tsilfidis and J Mourjopoulos, "Blind single-channel suppression of late reverberation based on perceptual reverberation modeling," *Journal of the Acoustical Society of America*, vol. 129(3), pp. 14391451, 2011.

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

[7] Uwe Zaeh, Korbinian Riedhammer, Tobias Bocklet, and Elmar Nöth, "Clap Your Hands! Calibrating Spectral Subtraction for Dereverberation," in *Proc, of the IEEE ICASSP*, 2010, pp. 4226–4229.

[8] 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*, Aalborg, Denmark, 2011.

[9] Rama Ratnam, Douglas L. Jones, Bruce C. Wheeler, Jr. William D. O'Brien, Charissa R. Lansing, and Albert S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[1] Audio demos can be found at `http://www.wcl.ece.upatras.gr/audiogroup/tools/derev.html`

[10] Dragana Sumarac-Pavlovic, Miomir Mijic, and Husnija Kurtovic, "A simple impulse sound source for measurements in room acoustics," *Applied Acoustics*, vol. 69, no. 4, pp. 378 – 383, 2008.

[11] Bruno H. Repp, "The sound of two hands clapping: An exploratory study," *The Journal of the Acoustical Society of America*, vol. 81, no. 4, pp. 1100–1109, 1987.

[12] M Woelfel, "Enhanced speech features by single-channel joint compensation of noise and reverberation," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 17, pp. 312–323, 2009.

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

[14] E. Georganti, J. Mourjopoulos, and F. Jacobsen, "Analysis of room transfer function and reverberant signal statistics," in *Proc. of the Acoustics '08*, Paris, France, June-July 2008.

[15] E. Georganti, T. Zarouchas, and J. Mourjopoulos, "Reverberation analysis via response and signal statistics," in *128th AES convention Proc.*, London, UK, 2010.

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

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

[18] Takayuki Hidaka, Yoshinari Yamada, and Takehiko Nakagawa, "A new definition of boundary point between early reflections and late reverberation in room impulse responses," *The Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 326–332, 2007.