Single–Channel Speech Dereverberation based on Spectral Subtraction

E.A.P. Habets
Technische Universiteit Eindhoven, Department of Electrical Engineering,
Signal Processing Systems Group, EH 3.27, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract— **Speech signals recorded with a distant microphone usually contain reverberation, which degrades the fidelity and intelligibility of speech in devices such as 'hands–free' conference telephones, automatic speech recognition and hearing aids. One important effect of reverberation on speech is overlap–masking, i.e. the energy of the previous phonemes is smeared over time, and overlaps following phonemes. In [1] a single–channel speech dereverberation method based on Spectral Subtraction was introduced to reduce this effect. The described method estimates the power spectrum of the reverberation based on a statistical model of late reverberation. This model depends on one parameter, the reverberation time. However, the reverberation time is frequency dependent due to frequency dependent reflection coefficients of walls and other objects and the frequency dependent absorption coefficient of air. In this paper, we have taken this dependency into account and studied the effect on reverberation reduction and distortion. The algorithm is tested using synthetically reverberated signals. The performances for different room impulse responses with reverberation times ranging from approximately 200 to 1200 ms show significant reverberation reduction with little signal distortion.**

Keywords— **dereverberation, spectral subtraction, speech enhancement.**

I. Introduction

In general, acoustic signals radiated within a room are linearly distorted by reflections from walls and other objects. These distortions degrade the fidelity and intelligibility of speech in devices such as 'hands–free' conference telephones, hearing aids and automatic speech recognition. Reverberation and spectral coloration cause users of hearing aids to complain of being unable to distinguish one voice from another in a crowded room. It is therefore of significant importance to investigate the application of signal processing techniques to the enhancement of speech or music distorted in an acoustic environment.

The problem of speech dereverberation has received a lot of attention from the seventies until now. Reverberation reduction processes may generally be divided into single or multiple microphone methods and into those primarily affecting coloration or those affecting reverberant tails. Early room echoes mainly contribute to coloration, or spectral distortion, while late echoes, or reverberation, contribute noise-like perceptions or tails on speech signals [2]. Looking at the signal processing techniques involved, the different approaches can be divided into three categories. Approaches in the first category, speech enhancement, explicitly exploit the characteristics of speech or the effect of reverberation on speech. Both single and multiple microphone techniques are exploited. Approaches in the second category, spatial processing, use multiple microphones placed at different locations. The multiple input signals can be manipulated to enhance or attenuate signals emanating from particular directions. Some methods in this category are inspired by the mechanisms of audition in the hearing system of animals and humans. The last category consists of approaches known as blind deconvolution. They are based upon channel identification to determine the Room Impulse Response (RIR) between the source and the receiver and use this information to equalize the channel. However, deconvolution methods have been shown to be little robust to small changes in the RIR.

In this paper we focus on an important effect of reverberation on speech which is referred to as overlap–masking, i.e. the energy of the previous phonemes is smeared over time, and overlaps following phonemes. In [1] a single–channel speech dereverberation method based on Spectral Subtraction was introduced to reduce this effect. The described method estimates the power spectrum of the reverberation based on a statistical model of late reverberation. This model depends on one parameter, which is directly related to the reverberation time of the room. However, the reverberation time is frequency dependent due to frequency dependent reflection coefficients of walls and other objects and the frequency dependent absorption coefficient of air. In this paper we have taken this dependency into account and studied the effect on reverberation reduction and distortion.

The outline of this paper is as follows. In Section II the statistical model for late reverberation is introduced. Section III describes the reverberant signal model. The Short Time Spectral Modification is described in Section IV. The complete algorithm and related implementation aspects are discussed in Section V. The performances for different reverberation times are discussed in Section VI, and finally conclusions and directions for future research are discussed in the last section.

II. Room Impulse Response Model

The model we use was developed by Polack in [3]. This model describes a Room Impulse Response (RIR) as one realization of a non-stationary stochastic process. A simplified version of this model can be expressed as

$$h(t) = \begin{cases} b(t)e^{-\Delta t} & t \geq 0 \\ 0 & t < 0 \end{cases}, \tag{1}$$

where $b(t)$ is a white zero–mean Gaussian stationary noise and Δ is linked to the reverberation time T_r through

$$\Delta \triangleq \frac{3\ln(10)}{T_r}.$$

The energy envelope of the RIR can be expressed as

$$E\{h^2(t)\} = \sigma^2 e^{-2\Delta t}, \tag{2}$$

where $E\{\cdot\}$ denotes ensemble averaging, and σ^2 denotes the variance of $b(t)$.

In contrast to the model in (1) the reverberation time is frequency dependent due to frequency dependent reflection coefficients of walls and other objects and the frequency dependent absorption coefficient of air [4]. This simple model can be extended to account for the frequency dependence of the reverberation time, by making Δ a function of frequency.

The RIR can be split into two components, $h_d(t)$ and $h_r(t)$ so that

$$h_d(t) = \begin{cases} h(t) & 0 \le t < T \\ 0 & otherwise \end{cases}$$

$$h_r(t) = \begin{cases} h(t) & t \ge T \\ 0 & otherwise \end{cases}$$

The value T is chosen such that $h_d(t)$ consists of the direct signal and a few early echoes and $h_r(t)$ consists of all later echoes, i.e. late reverberation. T usually ranges from 50 to 80 ms.

III. Reverberation Signal Model

The full-band reverberant signal results from the convolution of the anechoic speech signal $s(t)$ and the causal RIR $h(t)$:

$$x(t) = \int_{-\infty}^{t} s(\theta)h(t-\theta)\, d\theta.$$

The auto-correlation $r_{xx}(t, t+\tau)$, shortly denoted by r_{xx}, of the reverberant signal x at time t is

$$r_{xx} = \int_{-\infty}^{t}\int_{-\infty}^{t+\tau} E\{s(\theta)s(\theta')\}E\{h(t-\theta)h(t+\tau-\theta')\}\, d\theta d\theta'. \tag{3}$$

Using (1) it is possible to show that

$$E\{h(t-\theta)h(t+\tau-\theta')\} = e^{-2\Delta t}\sigma^2 e^{\Delta(\theta+\theta'-\tau)}\delta(\theta-\theta'+\tau),$$

where $\delta(\cdot)$ denotes the Dirac function. Equation (3) leads to

$$\begin{aligned} r_{xx} &= e^{-2\Delta t}\int_{-\infty}^{t} E\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\Delta\theta}\, d\theta \\ &= e^{-2\Delta t}\int_{-\infty}^{t-T} E\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\Delta\theta}\, d\theta \\ &\quad + e^{-2\Delta t}\int_{t-T}^{t} E\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\Delta\theta}\, d\theta. \end{aligned} \tag{4}$$

The auto-correlation at time t can be divided into two terms. The first term depends on the late reverberant signal and is responsible for overlap masking, whereas the second depends on the direct signal between time $t-T$ and t. Let us consider the auto-correlation at time $t-T$

$$r_{xx}(t-T, t-T+\tau) = e^{-2\Delta(t-T)}\int_{-\infty}^{t-T} E\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\Delta\theta}\, d\theta. \tag{5}$$

Using (4) and (5) we can now see that the auto-correlation at time t can be expressed as

$$r_{xx}(t, t+\tau) = r_{x_r x_r}(t, t+\tau) + r_{x_d x_d}(t, t+\tau),$$

with

$$r_{x_r x_r}(t, t+\tau) = e^{-2\Delta T} r_{xx}(t-T, t-T+\tau), \tag{6}$$

$$r_{x_d x_d}(t, t+\tau) = e^{-2\Delta t}\int_{t-T}^{t} E\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\Delta\theta}\, d\theta.$$

Using (1) it can also be shown that $E[x_d(t)x_r(t+\tau)] = 0$, where $x_d(t)$ and $x_r(t)$ result from the convolution of the anechoic speech signal $s(t)$ by respectively $h_d(t)$ and $h_r(t)$. In practice we can approximate the cross-correlation using a time average cross-correlation. Since $h(t)$ results from just one realization of the stochastic process for a fixed source and receiver position, this approximation will in general be unequal to zero. However, experiments have shown that the short-term time averaged cross-correlations are approximately zero. This results from the fact that the clean speech signal is quasi-stationary.

In practice the signals can be considered as stationary over periods of time that are short compared to the reverberation time T_r. This is justified by the fact that the exponential decay is very slow, and that speech is quasi-stationary. Let T_s be the time span over which the speech signal can be considered stationary, which is usually around 20-40 ms. We consider that $T_s \le T \ll T_r$. Under these assumptions, the counterparts of (5) and (6) in terms of the short-term power spectral densities are:

$$\begin{aligned} \gamma_{xx}(t, f) &= \gamma_{x_r x_r}(t, f) + \gamma_{x_d x_d}(t, f), \\ \gamma_{x_r x_r}(t, f) &= e^{-2\Delta(f)T}\gamma_{xx}(t-T, f). \end{aligned}$$

Since $x_d(t)$ and $x_r(t)$ are uncorrelated we may treat the late reverberant signal as an additive noise signal. The short-term Power Spectral Density (PSD) of the late reverberant signal can be estimated using a delayed and frequency dependent attenuated version of the short-term PSD of the reverberant signal. Therefore, we can estimate the short-term PSD of the direct signal by spectral subtraction of the late reverberant PSD.

IV. Short Time Spectral Modification

Numerous techniques for the enhancement of noisy speech degraded with uncorrelated additive noise have been proposed in literature. Among them the spectral subtraction methods are the most widely used due to the simplicity of implementation and the low computational load, which makes them the primary choice for real-time applications. A common feature of this technique is that the

noise reduction process can be related to the estimation of a Short-Time Spectral Attenuation factor. Since the spectral components are assumed to be statistical independent, this factor is adjusted individually as a function of the relative local *A Posteriori Signal to Noise Ratio* on each frequency. The A Posteriori SNR is defined as

$$SNR_{post}(t,f) \triangleq \frac{|X(t,f)|^2}{\gamma_{x_r x_r}(t,f)}, \tag{7}$$

where $\gamma_{x_r x_r}(t,f) = E[|X_r(t,f)|^2]$.

The Short-Time Spectral Attenuation factor can in general be defined as

$$G(t,f) = \left(1 - \left(\frac{1}{SNR_{post}(t,f)}\right)^{\beta_1}\right)^{\beta_2}. \tag{8}$$

Methods like Magnitude Subtraction ($\beta_1 = 1/2$, $\beta_2 = 1$), Power Subtraction ($\beta_1 = 1$, $\beta_2 = 1/2$) and Wiener Estimation ($\beta_1 = 1$, $\beta_2 = 1$) can be derived from (8). More complicated functions include non-linear gain functions or the Ephriam-Malah gain functions, which make a minimum mean square error estimate of the amplitude of the clean speech signal in the spectral or in the log-spectral domain. Other spectral subtraction techniques incorporate properties of the human auditory system.

Using informal listening tests we concluded, similar to the findings in [1], that Magnitude Subtraction gives beter performance compared to Power Subtraction and Wiener Estimation. This results in the following gain function

$$G(t,f) = 1 - \frac{1}{\sqrt{SNR_{post}(t,f)}}. \tag{9}$$

The estimate of the amplitude spectrum of the signal is given by

$$|\hat{S}(t,f)| = G(t,f)|X(t,f)|. \tag{10}$$

In all frames it is however possible that for some frequencies the estimated amplitude of the noise spectrum is larger than the instantaneous amplitude of the noisy speech spectrum $|X(t,f)|$. Since this could lead to negative estimates for the amplitude of the clean speech spectrum $|\hat{S}(t,f)|$, for these frequencies the gain function $G(t,f)$ is usually put to zero (i.e. half-wave rectification) or equal to a small noise floor value. However, because of the non-stationary character of the speech signal, this non-linear rectification mapping leads to a specific kind of residual noise, called musical noise, which consists of short-lived tones with randomly distributed frequencies. Different techniques have been proposed to eliminate this annoying residual noise, e.g. by averaging the (instantaneous) noisy speech spectrum over a number of frames, by augmenting the gain function with a soft-decision Voice Activity Detection or by using non-linear spectral subtraction techniques.

The residual noise problem is alleviated using two standard modifications. The first modification consists of averaging the *instantaneous SNR* (SNR_{inst}) in the calculation of the gain, yielding a reduction of the random variations due to the contribution of late reverberation in $|X(t,f)|$. The instantaneous SNR is defined as

$$SNR_{inst}(t,f) \triangleq SNR_{post}(t,f) - 1. \tag{11}$$

The averaged instantaneous SNR results in an estimate of the *a priori SNR* (SNR_{prio}) which is defined as

$$SNR_{prio}(t,f) \triangleq \frac{E[|X_d(t,f)|^2]}{E[|X_r(t,f)|^2]}.$$

In practice we can estimate SNR_{prio} using a moving average of SNR_{inst}

$$\begin{aligned} SNR_{prio}(t,f) &= \beta SNR_{prio}(t-1,f) \\ &\quad +(1-\beta)\mathrm{P}[SNR_{inst}(t,f)], \end{aligned} \tag{12}$$

where P denotes half-wave rectification and β ($0 \leq \beta \leq 1$) denotes the forgetting factor. The second modification consists in using a spectral floor as proposed in [5]. Instead of putting the negative of $|\hat{S}(t,f)|$ to zero, the values of $|\hat{S}(t,f)|$ less then a threshold, equal to $\lambda|X(t,f)|$ are set to this threshold. Applying above modifications to the standard gain function in (9) results in the following gain function

$$G(t,f) = \begin{cases} 1 - \frac{1}{\sqrt{SNR_{prio}(t,f)+1}} & \text{if } |\hat{S}(t,f)| \geq \lambda|X(t,f)| \\ \lambda & \text{otherwise} \end{cases} \tag{13}$$

V. Implementation

The signals are digitized with a sampling rate of 8 kHz. In the following, the discrete time and frame indices will be denoted by n and m, respectively, and the discrete frequency index k. An overview of the complete algorithm is presented in Figure 1.

The different stages of the algorithm can be described as follows:

Time Frequency Analysis and Synthesis. The Time Frequency (TF) analysis can be performed in many ways. As an example we used the Short Time Fourier Transform. Although this analysis results in a constant time-frequency bandwidth product, it performs well and has a low computational complexity. The analysis window is a 128 point hamming window, and the overlap between two successive windows is set to 75%. Each frame is zero padded to 256 points in order to avoid wrap around errors. The estimated dereverberated signal $\hat{s}(n)$ is then reconstructed through the overlap-add technique [6] from the estimated amplitude spectrum $|\hat{S}(m,k)|$ and the reverberant phase signal.

Blind Estimation of $T_r(k)$ and $\gamma_{x_r x_r}(m,k)$. In order to estimate the late reverberant PSD we need to estimate the reverberation time of the room. Partially blind methods have been developed in which the characteristics of the room are 'learned' using neural network approaches [7], or some form of segmentation procedure is used for detecting gaps in sounds to allow the sound decay curve to be

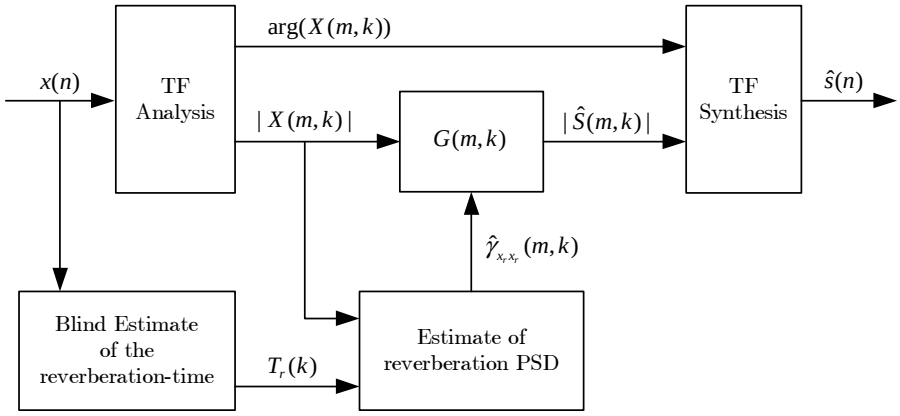

Fig. 1. Overview of the algorithm.

tracked [1]. Recently, a blind method has been proposed by R. Ratnam et.al. based on a maximum-likelihood estimation procedure [8]. These methods can also be applied to sub-band signals in order to estimate the frequency dependent reverberation time. For evaluation purposes we have determined the reverberation time directly from the RIRs and the filtered RIRs using Schroeder's method [9]. The filtered RIRs are constructed using 6 one-octave bandpass filters. The center frequencies (in Hz) of these bands are $f_c = [125, 250, 500, 1000, 2000, 4000]$.

The short-term PSD $\gamma_{x_r x_r}(m,k)$ is estimated by

$$\begin{aligned} \hat{\gamma}_{xx}(m,k) &= \beta_x \hat{\gamma}_{xx}(m-1,k) + (1-\beta_x)|X(m,k)|^2, \\ \hat{\gamma}_{x_r x_r}(m,k) &= e^{-2\Delta(k)T} \hat{\gamma}_{xx}(m-T',k), \end{aligned}$$

with $\beta_x = 0.9$ and $T' = \left\lfloor \frac{T f_s}{32} \right\rfloor$.

Magnitude Spectral Subtraction. The discrete versions of (7), (11), (12), (13) and (10) are used to obtain the estimate $|\hat{S}(m,k)|$. The forgetting factor β in (12) was set to 0.5. The threshold λ in (13) was set to 0.1, corresponding to a maximum attenuation of 20 dB.

VI. Experimental Results

The reverberant microphone signals are obtained by convolution of an anechoic female voice of 4 seconds by different synthetic RIRs. The dimensions of the room are 5 m x 6 m x 4 m (l x b x h). The RIRs are constructed using Allen and Berkley's image method [10]. The source-receiver distance equals 3.5 m. The measured full-band and sub-band reverberation times (in ms) are presented in Table I. The parameter T was set to 60 ms.

A. Objective Measurements

Objective measurements for both the reverberation reduction and the distortion have been used for this evaluation. The reverberant signal is decomposed into the sum of a direct signal $d_{in}(n)$ and a reverberant part $r_{in}(n)$, obtained by convolving the anechoic signal with the first 6 ms (w.r.t. the direct sound) of the RIR, and with the RIR minus this part. While the complete reverberant signal is being processed, the time-varying, signal dependent gain function is recorded. The recorded gain is then applied to the reverberant part, which results in $r_{out}(n)$. The input and output signals are filtered using 6 one-octave bandpass filters, the center frequencies are equal to those described in Section V.

RIR	*Sub-Band*						*Full-Band*
	1	2	3	4	5	6	
1	300	235	197	177	214	190	194
2	665	345	309	290	336	336	329
3	1230	523	404	378	430	431	427

TABLE I
Reverberation time $T_r(b)$ (ms).

Reverberation Reduction When no speech is present in the anechoic signal the global reverberation reduction is calculated using

$$RR(b) = 10 \log_{10} \left(\frac{\sum_{n \in \Omega_{\text{Silence}}} r_{in}^2(n,b)}{\sum_{n \in \Omega_{\text{Silence}}} r_{out}^2(n,b)} \right) \quad \text{(dB)}$$

where $b \in \{1, \ldots, 6\}$ denotes the sub-band index. The separation between speech and silence zones has been made trough manual segmentation.

Distortion The log spectral distance (LSD) between the direct signal and the output signal is used as a measure of distortion. The distance is calculated in each sub-band using the following expression

$$LSD(m,b) = \frac{1}{f_\Delta(b)} \sum_{k \in \mathcal{B}(b)} \left| \log \left(\frac{|\hat{S}(m,k)|^2}{|D_{in}(m,k)|^2} \right) \right|,$$

where $f_\Delta(b)$ denotes the bandwidth of sub-band b and $\mathcal{B}(b)$ consist of a set containing all discrete frequencies in sub-band b. Finally, the mean log spectral distance over the periods of speech is calculated for each sub-band.

B. Results

The global reverberation reduction and mean log spectral distance for the three RIRs are shown in Figure 2 and 3.

The results obtained using the full-band (FB) and sub-band (SB) reverberation times are depicted using a star (*) and a diamond ($\diamond$), respectively. In those sub-bands where the full-band reverberation time is larger than the sub-band reverberation time, the distortion is reduced by using the sub-band reverberation times while the reverberation reduction is comparable. In those sub-bands where the full-band reverberation time is smaller then the sub-band reverberation time, the reverberation reduction is increased by using the sub-band reverberation time. However, in the low frequency sub-bands we notice an increase in the log spectral distance. This is mainly caused by the fact that Polack's stochastic model is valid only above the Schroeder frequency [3], which is defined as $f_{schroeder} = 2000\sqrt{\frac{T_r(f)}{V}}$ (Hz), where V (m^3) is the volume of the room.

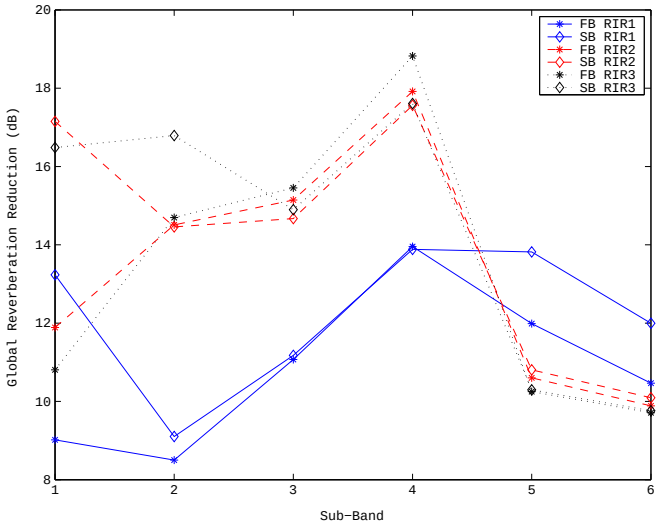

Fig. 2. Global Reverberation Reduction for different RIRs.

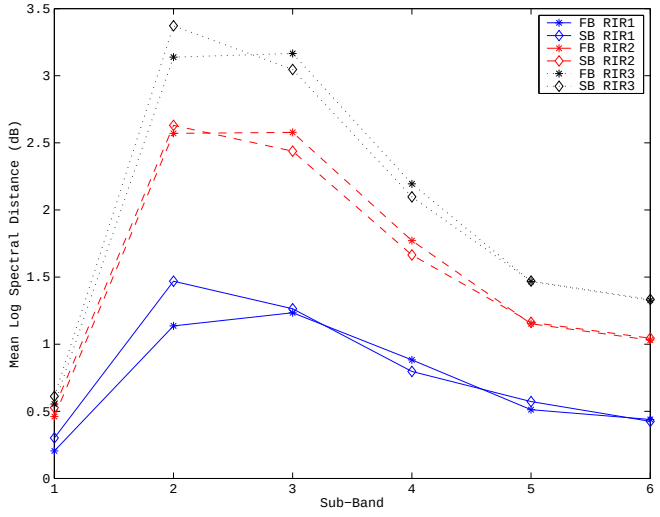

Fig. 3. Mean Log Spectral Distance for different RIRs.

VII. Conclusions

In this paper we have described a single-channel speech dereverberation algorithm based on magnitude spectral subtraction. A stochastic model has been used to model the exponential decay of late reverberation. This model depends on one parameter which is directly related to the reverberation time. The short-term Power Spectral Density (PSD) of the late reverberant signal is estimated using this model. It can be subtracted from the short-term PSD of the reverberant signal, yielding an estimate of the direct signal. By taking the frequency dependency of the reverberation time into account we have shown that depending on the reverberation characteristics of the room the performance, in terms of speech distortion and reverberation reduction, is increased.

Results have shown that over-estimation of the reverberation time for a specific sub-band results in unnecessary speech distortions in this sub-band, these distortions can be reduced using a more accurate, i.e. frequency dependent, reverberation time. Under-estimation of the reverberation time for a specific sub-band results in low reverberation reduction in this sub-band. In this case, exploiting the frequency dependent reverberation time results in higher reverberation reduction. Unfortunately this also results in an increase in speech distortion for low frequencies, which is mainly caused by the fact that Polack's stochastic model is only valid above the Schroeder frequency.

VIII. Acknowledgements

The author expresses his thanks to STW (project EEL 4921) for funding and would like to thank Ir. J. v.d. Laar and his supervisor Dr. Ir. P.C.W. Sommen very much for carefully proofreading the manuscript.

References

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

[2] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *Journal of the acoustical society of america*, vol. 62, no. 4, pp. 912–915, 1977.

[3] J.D. Polack, *La transmission de l'énergie sonore dans les salles*, Thèse de doctorat d'etat, Université du Maine, La mans, 1988.

[4] H. Kuttruff, *Room Acoustics*, Spon Press, London, fourth edition, 2000.

[5] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE ICASSP'79*, vol. 4, pp. 208–211, 1979.

[6] J. Allen and L. Radiner, "A unified approach to short-time fourier analysis and synthesis," *IEEE Proceedings*, vol. 65, 1977.

[7] T. J. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *Journal of the audio engineering society*, vol. 49, pp. 219230, 2001.

[8] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. OBrien Jr., C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *Journal of the acoustical society of america*, vol. 114, no. 5, pp. 28772892, November 2003.

[9] M. R. Schroeder, "New method of measuring reverberation time," *Journal of the acoustical society of america*, vol. 37, pp. 409, 1965.

[10] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *journal of the acoustical society of america*, vol. 65, no. 4, pp. 943–950, 1979.