acta acustica · ACUSTICA
Vol. 87 (2001) 359 – 366



# A New Method Based on Spectral Subtraction for Speech Dereverberation

K. Lebart, J. M. Boucher
ENST-Br, Dept. SC., Technopôle de Brest Iroise, BP 832, 29285 Brest Cedex, France
E-mail: lebart@cee.hw.ac.uk, JM.Boucher@enst-bretagne.fr

P. N. Denbigh
University of Sussex, Falmer, Biomedical Engineering, BN1 9QT, Brighton, England

**Summary**
A new monaural method for the suppression of late room reverberation from speech signals, based on spectral subtraction, is presented. The problem of reverberation suppression differs from classical speech de-noising in that the "reverberation noise" is non stationary. In this paper, the use of a novel estimator of the non-stationary reverberation-noise power spectrum, based on a statistical model of late reverberation, is presented. The algorithm is tested on real reverberated signals. The performances for different RIRs with $T_r$ ranging from 0.34 s to 1.7 s consistently show significant noise reduction with little signal distortion. Moreover, when used as a front end to an automatic speech recognition system, the algorithm brings about dramatic improvements in terms of automatic speech recognition scores in various reverberant environments.

**PACS no. 43.00.Xx, 00.00.Xx**

## 1. Introduction

Reverberation is an acoustical noise appearing in enclosed spaces through the multiple reflections and diffractions of the sound on the walls and objects of a room. When a speaker talks in a room, these multiple echoes add to the direct sound and blur its temporal and spectral characteristics. Its effect can be alleviated by the use of a microphone close to the source of the signal of interest. However, this is not convenient for "hand free" applications, such as for instance men-machine communication. Indeed many applications for which a distant sound pick up is required perform poorly in the presence of reverberation. This is the case for Automatic Speech Recognition (eg. [1]) or Automatic Speaker Verification (eg. [2]). Dereverberation can also be of benefit to hearing impaired listeners since reverberation can reduce speech intelligibility [3].

The problem of speech dereverberation has received a lot of attention from the seventies until now. The process of reverberation can be modeled as a filtering process: the speech signal is convolved by the impulse response of the acoustic channel defined by the emitter, the receiver and the surrounding environment. Such an impulse response is referred to as a Room Impulse Response (RIR) [1]

A first set of methods rely on this model and aims at deconvolving the reverberated speech signal ([4, 5, 6, 7]). However, deconvolution methods require the RIR to be known precisely, and have been shown to be little robust to small changes in the RIR ([8, 9]). In the applications considered here, the RIR is unknown and varying. Techniques such as sub-band envelope deconvolution ([10, 11, 12]) or envelope expansion methods are more robust to RIR variation. They aim at increasing the modulation depth of the reverberated speech. They have been suggested to tackle both noise [13] and reverberation [14].

Another set of methods use the spatial and directional properties of the reverberation noise, considered to be an additive noise. Array processing techniques have been proposed (eg. [15, 16]). Methods inspired by the mechanisms of audition in the hearing system of animals -and humans- have been suggested ([17, 18, 19, 20, 21, 22]) along with more classical array processing methods ([23]). A group of algorithms that use the spatial decorrelation of late reverberation stemmed from the work of Allen, Berkley and Blauert [24]: [25, 26, 27, 28, 29, 8, 30].

In this paper, we focus on an important effect of reverberation on speech which is referred to as overlap-masking [31, 32]: the energy of previous phonemes is smeared over time, and overlaps following phonemes. This results in the blur and masking of the spectral features of the phonemes.

The actual physical process underlying this smearing is the multiple reflections and diffusions of the sound waves on the boundaries and obstacles of the room, corresponding to late reverberation. As a result of the phenomena of absorption by the air and the reflectors, the reverberated energy decays exponentially, with a time constant depending on the characteristics of the room.

Intuitively, the evolution along time of the energy of the reverberant tails for a phoneme will have an exponential decay behaviour similar to that of the Room Impulse Response (RIR). The repartition along frequency of the reverberant energy will depend on the repartition along frequency of the energy of the excitation, that is the spectrum of the considered phoneme.

Received 15 December 1999,
accepted 2 April 2001.

[1] The name "Room Impulse Response" can be misleading, since it is not the room that a RIR caracterizes, but rather a specific acoustic channel within this room.

It therefore seems that the smearing of the energy of the speech signal into reverberation tails can be coarsely modeled from the knowledge of the preceding phonemes and of the reverberation time of the room. This modeling can in turn be used to estimate and suppress part of the reverberant energy from the reverberant speech signal.

The following study will try to formalize these ideas by using of a statistical model of late reverberation. This model is detailed in section 2 and leads to an equation linking the power spectral density (PSD) of the reverberation part of the signal to that of the reverberated signal. Section 3 then details the dereverberation algorithm based on this model. In section 4 the performance of the algorithm is assessed for different situations. Section 5 presents a discussion on possible improvements for the algorithm.

## 2. Model

### 2.1. Model for the Room Impulse Response

The Room Impulse Response is modeled as the outcome of a non-stationary random process:

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

where $b(t)$ is a zero-mean Gaussian stationary noise, considered in first approximation to be white, and $\Delta$ is linked to the reverberation time $T_r$ through:

$$\Delta = \frac{3\ln 10}{T_r}.$$

This model was proposed by Polack [33], after Moorer [34], for application to artificial reverberation.

### 2.2. Model for the Reverberant Signal

Let us consider $s(t)$ to be the anechoic speech signal, and $x(t)$ to be the reverberated speech signal, resulting from the convolution of $s(t)$ by the RIR $h(t)$:

$$\begin{aligned} x(t) &= \int_{-\infty}^{+\infty} s(\theta)h(t-\theta)\,\mathrm{d}\theta \\ &= \mathrm{e}^{-\Delta t}\int_{-\infty}^{t} s(\theta)b(t-\theta)\mathrm{e}^{\Delta\theta}\,\mathrm{d}\theta \end{aligned} \tag{2}$$

since $h(t)$ is causal. Then, if $s$ and $b$ are considered to be independent random processes, the autocorrelation of $x$ at time t is:

$$\begin{aligned} &E[x(t)x(t+\tau)] = \\ &\mathrm{e}^{-2\Delta t}\int_{-\infty}^{t}\int_{-\infty}^{t+\tau} E\big[s(\theta)s(\theta')\big]E\big[b(t-\theta)b(t+\tau-\theta')\big] \\ &\qquad\cdot \mathrm{e}^{\Delta(\theta+\theta'-\tau)}\,\mathrm{d}\theta\,\mathrm{d}\theta'. \end{aligned} \tag{3}$$

Since $b$ is considered to be a white noise of power $\sigma_0^2$:

$$E\big[b(t-\theta)b(t+\tau-\theta')\big] = \sigma_0^2\delta(\theta-\theta'+\tau),$$

where $\delta()$ represents the Dirac function.

Equation (3) leads to:

$$\begin{aligned} &E\big[x(t)x(t+\tau)\big] = \\ &\mathrm{e}^{-2\Delta t}\int_{-\infty}^{t} E\big[s(\theta)s(\theta+\tau)\big]\sigma_0^2 e^{2\Delta\theta}\,\mathrm{d}\theta. \end{aligned}$$

Let us now consider the autocorrelation of $x$ at a later time $t+T$: $A = E[x(t+T)x(t+T+\tau)]$:

$$\begin{aligned} A &= \mathrm{e}^{-2\Delta(t+T)}\int_{-\infty}^{t+T} E\big[s(\theta)s(\theta+\tau)\big]\sigma_0^2\mathrm{e}^{2\Delta\theta}\,\mathrm{d}\theta \\ &= \mathrm{e}^{-2\Delta T}\mathrm{e}^{-2\Delta t}\int_{-\infty}^{t} E\big[s(\theta)s(\theta+\tau)\big]\sigma_0^2\mathrm{e}^{2\Delta\theta}\,\mathrm{d}\theta \\ &\quad + \mathrm{e}^{-2\Delta(T+t)}\int_{t}^{t+T} E\big[s(\theta)s(\theta+\tau)\big]\sigma_0^2\mathrm{e}^{2\Delta\theta}\,\mathrm{d}\theta. \end{aligned} \tag{4}$$

This equation can have different interpretations. They are detailed in the next paragraph.

### 2.3. Interpretations

From equation (4), it can be seen that:

$$\begin{aligned} &E\big[x(t+T)x(t+T+\tau)\big] = \\ &\qquad \mathrm{e}^{-2\Delta T}E\big[x(t)x(t+\tau)\big] \\ &\quad + \mathrm{e}^{-2\Delta(T+t)}\int_{t}^{t+T} E\big[s(\theta)s(\theta+\tau)\big]\sigma_0^2\mathrm{e}^{2\Delta\theta}\,\mathrm{d}\theta. \end{aligned} \tag{5}$$

The autocorrelation of $x$ at time $t+T$ is the sum of two terms. The first term depends on the past reverberated signal, whereas the second depends on the anechoic signal between time $t$ and $t+T$. The first term is considered as being responsible for overlap masking, since its energy over the time interval $[t,\ t+T]$ is entirely due to the reverberated signal present at times prior to $t$.

Another interpretation of the two terms in equation (4) is possible: let $h(t)$ be split into two components, $h_d(t)$ and $h_r(t)$, so that:

$$h_d(t) = \begin{cases} h(t) \text{ if } 0 \leq t < T, \\ 0 \text{ otherwise,} \end{cases}$$

$$h_r(t) = \begin{cases} h(t) \text{ if } \quad t \geq T, \\ 0 \text{ otherwise.} \end{cases}$$

Let $s_d(t)$ and $r(t)$ be the results of the convolution of $s(t)$ by respectively $h_d(t)$ and $h_r(t)$. If $T$ is relatively much smaller compared to $T_r$, $s_d(t)$ is made up of the direct signal and a few early echoes. As a first approximation it can be considered as being the direct signal, whereas $r(t)$ corresponds to all the later echoes, that is to late reverberation. It can be shown [35] that the first term in equation (4) equals $E[r(t)r(t+\tau)]$ and the second term equals $E[s_d(t)s_d(t+\tau)]$. Equation (4) can therefore equivalently be written:

$$\begin{aligned} E\big[x(t)x(t+\tau)\big] &= E\big[r(t)r(t+\tau)\big] \\ &\quad + E\big[s_d(t)s_d(t+\tau)\big], \end{aligned} \tag{6}$$

with:

$$E\big[r(t)r(t+\tau)\big] = \mathrm{e}^{-2\Delta T} \cdot E\big[x(t-T)x(t-T+\tau)\big]. \quad (7)$$

In practice the signals can be considered as stationary over periods of time that are short compared to $T_r$. This is justified by the fact that the exponential decay is very slow, and that speech is quasi-stationary. Let $D$ be the typical time span over which the signal can be considered stationary. We consider that $D \leq T << T_r$. In practice, the order of magnitude of $D$ is about 50 ms and that of the reverberation times considered is around 1 s. Under these approximations, the counterpart of Eqs. (6 and 7) in terms of short-term power spectral densities are:

$$\gamma_{xx}(t,f) = \gamma_{rr}(t,f) + \gamma_{s_d s_d}(t,f), \quad (8)$$

$$\gamma_{rr}(t,f) = \mathrm{e}^{-2\Delta T}\gamma_{xx}(t-T,f). \quad (9)$$

## 3. Algorithm

### 3.1. Overview

The overview of the algorithm is presented in Figure 1.

The signals are digitized with a sampling rate of $8kHz$. In the following, the discrete time indexes will be notated $n$ or $m$, and the discrete frequency index $k$.

The reverberated signal $x(n)$ is decomposed into a Short-Time Fourier Transform (STFT) filter bank. 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 power spectral density of the reverberation noise $\hat{\gamma}_{rr}$ is estimated according to equation (9), as detailed in section 3.3. The square root of this estimate is then subtracted from the amplitude spectrum of the reverberated signal, $|X(m,k)|$, yielding an estimate of the amplitude spectrum of the dereverberated signal, $|\hat{S}(m,k)|$. In practice, this is realized by a short-term spectral attenuation, equivalent to spectral subtraction. This modification is detailed in section 3.2. The estimated dereverberated signal $\hat{s}(n)$ is then reconstructed from its estimated amplitude spectrum and the noisy phase, through the overlap-add technique (eg. [36]).

### 3.2. Short-Term Spectral Modification

A formulation for amplitude spectral subtraction is:

$$|\hat{S}(m,k)| = |X(m,k)| - \hat{\gamma}_{rr}^{1/2}(m,k) = G(m,k)|X(m,k)|,$$

with $m$ being the sub-band time index, $k$ being the frequency index, and

$$G(m,k) = \frac{|X(m,k)| - \hat{\gamma}_{rr}^{1/2}(m,k)}{|X(m,k)|} = 1 - \frac{1}{\sqrt{SNR_{pos}+1}}, \quad (10)$$

where $SNR_{pos} = \frac{|X|^2}{\hat{\gamma}_{rr}} - 1$, $|\hat{S}(m,k)|$ is an estimate of the amplitude spectrum of the signal, and $\hat{\gamma}_{rr}$ is an estimate of the average PSD of the noise.

In a comparative study of different short time spectral attenuation algorithms, Ayad [37] concludes that amplitude subtraction gives very good performance, compared with other more sophisticated methods. This method is the one retained in this article.

One of the problems arising from such implementation is that in practice, the term $|X(m,k)| - \hat{\gamma}_{rr}^{\frac{1}{2}}(m,k)$ can have negative values. This is due to the fact that $\hat{\gamma}_{rr}(m,k)$ is an estimate of the average noise spectrum. But the noise component in $|X(m,k)|$ can be inferior to the average. This leads to negative values for the estimate $|\hat{S}(m,k)|$ when no or little signal energy is present in the considered frame. To make up for this problem, a commonly used solution is to set to 0 the negative values of $|X(m,k)| - \hat{\gamma}_{rr}^{\frac{1}{2}}(m,k)$. However, this nonlinear rectification yields a specific residual noise, often referred to as "musical noise" to account for its perceptual characteristics.

Whenever the signal is present, musical noise is masked, but it is clearly perceptible during periods of silence. As a matter of fact, at times when the noise only is present, some frequency bands of $X(m,k)$ contain more energy than the average $\hat{\gamma}_{rr}(m,k)$. The effect of the spectral subtraction is to set all the other frequency bands to 0, while only attenuating those bands with more energy. The spectrum of the processed signal therefore contains peaks positioned randomly at isolated frequencies, lasting for an average duration of the length of the analysis window [38].

Many solutions have been proposed in the literature to tackle the problem of musical noise (see eg. [39, 40]...). In this article, two standard modifications are added to the algorithm to alleviate the problem of musical noise. The first one consists in averaging the term $SNR_{pos}$ in the calculation of the gain, yielding a reduction of the random variations due to the noise contribution in $|X(m,k)|$. The second one consists in using a spectral floor, as proposed in [41].

**Smoothing**

$G(m,k)$ in equation (10) can be substituted by:

$$G(m,k) \simeq 1 - \frac{1}{\sqrt{SNR_{pri}+1}}. \quad (11)$$

The term $SNR_{pri}$ in equation (11) is defined as:

$$SNR_{pri} = E\big[\frac{|X|^2}{\hat{\gamma}_{rr}} - 1\big]. \quad (12)$$

It is estimated through a running average:

$$SNR_{pri}(m,k) = \beta SNR_{pri}(m-1,k) + (1-\beta)\max\big[0, SNR_{pos}\big]. \quad (13)$$

The operator $\max[0,.]$ prevents the inclusion of negative values of $SNR_{pos}$ which have no physical meaning.

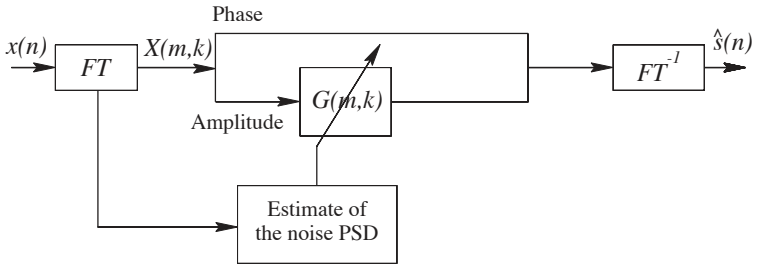


Figure 1. Overview of the algorithm.

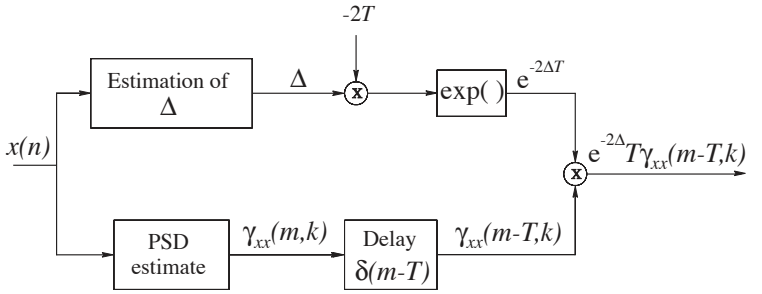


Figure 2. Estimation of the reverberant PSD.

**Spectral Floor**

Instead of putting the negative estimates of $|\hat{S}(m,k)|$ to 0, the values of $|\hat{S}(m,k)|$ less than a threshold, equal to $\lambda\sqrt{\hat{\gamma}_{rr}(m,k)}$, are set to this threshold. In practice $\lambda = 0.1$, corresponding to an attenuation of 20 dB, is used.

$$|\hat{S}(m,k)| = \begin{cases} G(m,k)|X(m,k)| \\ \qquad \text{when } \geq \lambda\sqrt{\hat{\gamma}_{rr}(m,k)}, \\ \lambda\sqrt{\hat{\gamma}_{rr}(m,k)} \quad \text{otherwise.} \end{cases}$$

The gain used in the algorithm is finally:

$$G(m,k) = \begin{cases} 1 - \frac{1}{\sqrt{SRN_{pri}+1}} \\ \qquad \text{if } |\hat{S}(m,k)| \geq \lambda\sqrt{\hat{\gamma}_{rr}(m,k)}, \\ \lambda\sqrt{\hat{\gamma}_{rr}(m,k)}/|X(m,k)| \qquad \text{otherwise.} \end{cases}$$

### 3.3. Estimation of the Reverberant PSD

From equation (9), it can be seen that two terms need to be estimated to obtain an estimate of the reverberation PSD, as shown in Figure 2: the parameter of the model $\Delta$ (or equivalently the reverberation time $T_r$) and the PSD of the past reverberated signal.

Then

$$\hat{\gamma}_{rr}(m,k) = \mathrm{e}^{-2\Delta T}\hat{\gamma}_{xx}(m-T,k).$$

Since the duration of stationarity of the signals is assumed to be about 20 ms, according to the approximation stated in paragraph 2.3, $T$ is set to: $T \simeq 50$ ms.

Estimation of the PSD of the past reverberated signal

The PSD of the past reverberated signal is estimated by averaging the periodograms of the signal. This is done by a running average according to:

$$\hat{\gamma}_{xx}(m,k) = \beta\hat{\gamma}_{xx}(m-1,k) + (1-\beta)|X(m,k)|^2. \quad (14)$$

If $\beta$ is close to 1, the variance of the estimate of the PSD is small, but the equivalent averaging duration long. The averaging time should be kept small since the signal is non stationary. $\beta$ should be chosen as a compromise so that the variance of the estimate is as small as possible while the assumption of quasi-stationarity is respected. In practice, $\beta$ was set to 0.9.

Estimation of $T_r$

The reverberation time is a characteristic of the room. It can change if the environment of the system changes, but its variations can be considered as very slow. It therefore only needs to be estimated from time to time, with the assumption that it does not change between the update periods. The estimation of $T_r$ comprises two distinct stages:

**Detection of silences** , where the energy of the reverberated signal decays exponentially. This is realized in two steps: first, the zones where the smoothed energy envelope of the signal is decreasing are automatically detected. Amongst these zones, only the longest ones are selected: they are deemed to correspond to silences in the speech signal, and therefore present the exponential decay corresponding to the reverberation time of the room.

**Estimation of $T_r$ over the silences** . The slope of the logarithm of the smoothed energy envelope over the decreasing periods is estimated through linear regression.

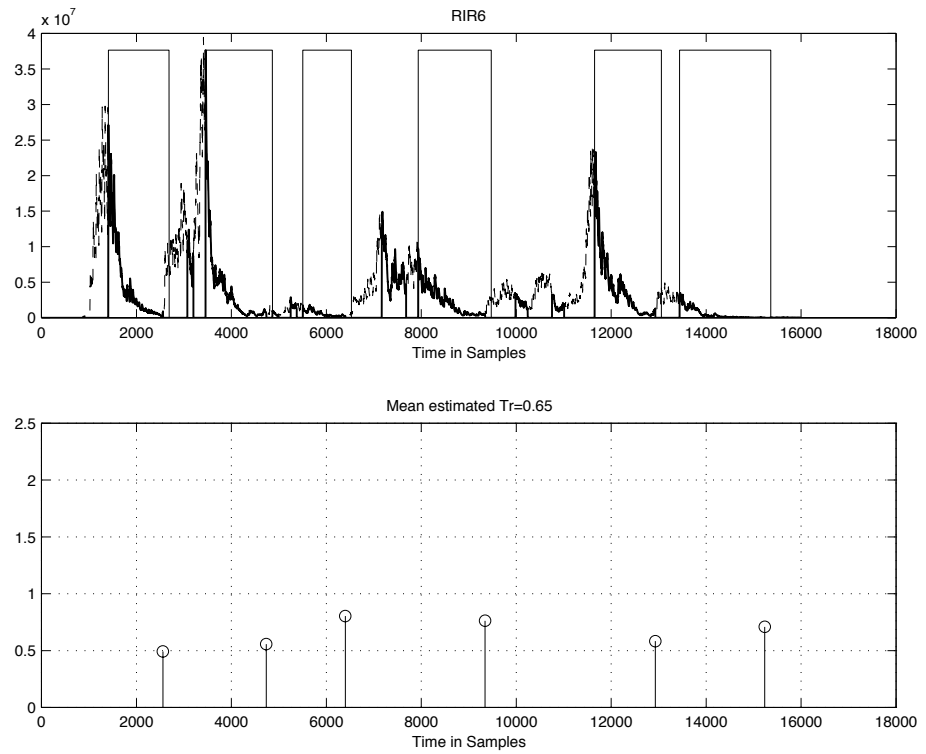


Figure 3. On-line estimation of $T_r$ for RIR6; Real value: 0.55 s.

In Figure 3, an example of the running estimation of $T_r$ over a reverberated phrase is presented.

In the upper panel, the smoothed energy envelope of a reverberated speech sentence is represented, along with the automatically selected 'exponential decay'-zones. The lower panel shows the values of $T_r$ estimated over these areas. The estimated values have been found to be in good accordance with the values of $T_r$ estimated on the Room Impulse Responses by the method proposed by Schroeder in [42].

## 4. Results

### 4.1. Methods of Assessment

The reverberated speech signals used were obtained by convolution of anechoic phrases by real room impulses (RIRs), measured by two closely spaced omni directional microphones on a dummy head. Six different RIRs were used, having reverberation times ranging from 0.4 s to 1.7 s. RIR3, RIR5 and RIR6 correspond to different acoustic channels in the same room. To assess the efficiency of the algorithm we have used four types of objective measurement:

**Input to Output SNR gain [8]:** We used the time varying method proposed in [8]. The reverberated signal is decomposed into a sum of a direct signal $s_{in}$ and a reverberant part $r_{in}$, obtained by convolving the anechoic signal with the first 5 ms of the RIR, and with the RIR minus its first 5 ms. While the complete reverberated signal is being processed, the time varying, signal-dependent gain is recorded. The recorded gain is then applied separately to the direct signal and reverberant part, giving respectively $s_{out}$ and $r_{out}$. The SNR gain is then defined as:

$$G_{SNR} = 10\log_{10}\left(\frac{\sum_{\text{Frame}} s_{out}^2 \sum_{\text{Frame}} r_{in}^2}{\sum_{\text{Frame}} s_{in}^2 \sum_{\text{Frame}} r_{out}^2}\right). \qquad (15)$$

It is calculated globally over the periods of speech activity.

**Noise Reduction:** When no speech energy is present in a frame, the noise reduction is calculated in the same way by:

$$NR = 10\log_{10}\left(\frac{\sum_{\text{Frame}} r_{in}^2}{\sum_{\text{Frame}} r_{out}^2}\right). \qquad (16)$$

The separation between speech and silence zones has been made through manual segmentation.

**Distortion:** A cepstral distance (CD) [43] between the direct signal at the input and the output of the system is used as a measure of distortion. Only the first 8 cepstral coefficients, which are linked to the first LPC coefficients, are taken into account. The distance used therefore reflects the dissimilarity in term of the formant structure of the two signals. This measure, along with the SNR gain, reflects how the speech quality is affected by the algorithm.

**Speech Recognition Scores:** A commercially available isolated-words speech recognizer is first trained in

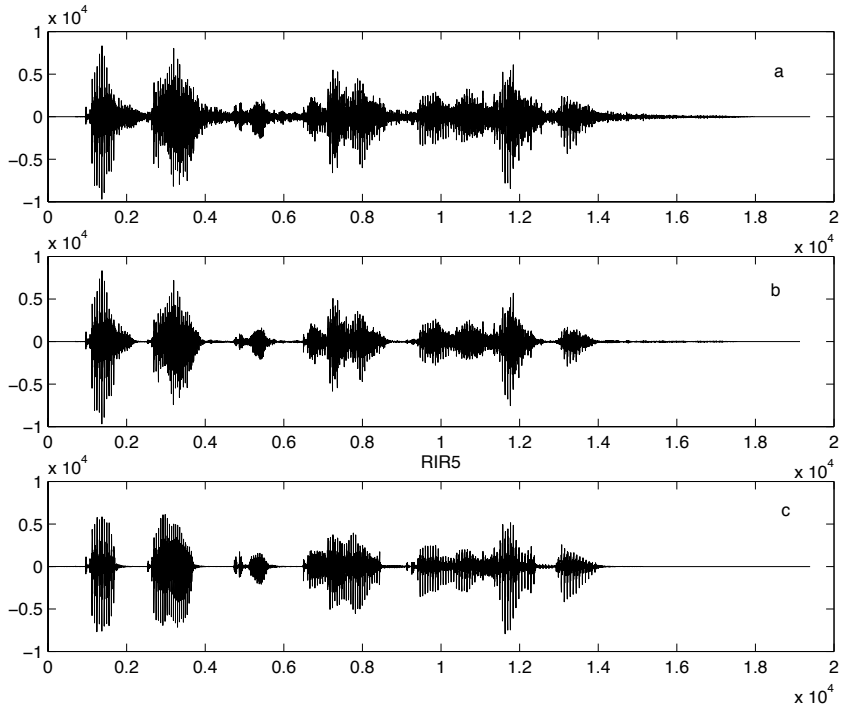


Figure 4. Reverberated (a), Processed (b), and anechoic (c) signals for RIR5.

anechoic conditions on a set of 240 isolated English words uttered by a male speaker, so that it achieves 99% recognition. The recognition score is then measured on the same set of words, artificially reverberated by convolution by one of the RIR. Since this does not correspond to the training conditions, the recognition rate drops dramatically. The recognition scores are then measured on the processed reverberated signal. The difference between the two last measures reflects the benefits of the algorithm.

### 4.2. Performance

Each measure of performance is estimated both for the presented algorithm (referred to as "Monaural" algorithm) and a reference algorithm, referred to as "Bloom" algorithm [44].

#### 4.2.1. Objective Measurements

The lower panel of Figure 4 shows the waveform of the French sentence *'Papa coupe l'herbe dans le jardin'*, in anechoic conditions. The upper panel shows the same sentence when reverberation is added. The middle panel shows the signal after is has been processed by our algorithm. The attenuation of the reverberant tails is striking.

The performance of the Monaural algorithm is presented in Table II and III. Table II shows the results for the algorithm without the on-line estimation of $T_r$ (the true value of $T_r$ is fed into the algorithm).

For comparison, the performance of the "Bloom" algorithm, in exactly the same conditions, are presented in Table I.

Table III shows the results for the algorithm including the on-line estimation of $T_r$.

The performance of the algorithm is not significantly different whether $T_r$ is set to its true value or to the values obtained from the on-line estimation. This shows that the influence of the value of $T_r$ is not critical, and the on-line estimation provides sufficiently accurate estimates of $T_r$ for the purpose of the algorithm. Indeed if $T_r$ is overestimated, this is equivalent to overestimating the average noise PSD, which is classically done (eg. [41, 38]) in order to suppress more noise.

The main improvement brought about by the algorithm is the noise reduction over silence periods. In most cases, the distortion of the signal remains fairly low. Informal listening has shown that some musical noise remains after processing, especially for the two longest RIR. This can be reduced further by adjusting the parameters of the algorithm, to the detriment of the noise reduction performance.

#### 4.2.2. Automatic Speech Recognition Scores

Automatic speech recognition scores for the reverberated signals processed by the algorithm are presented in Table IV.

The algorithm used for these tests is the one where $T_r$ is not estimated on-line, but fed into the algorithm. The results are compared to the ones obtained with a classical algorithm based on the spatial decorrelation of late reverberation: the "Bloom algorithm".

For the two longer room impulse responses, RIR1 and RIR2, the monaural algorithm proposed outperforms the classical Bloom algorithm by 30 points. RIR3 and RIR6 are somewhat milder in that they have a shorter reverberation time, but still a poor direct to reverberant ratio. Here, the Monaural algorithm yields recognition performance superior by 15 points. For RIR4, which corresponds to the

smaller $T_r$, and RIR5, which corresponds to the largest direct to reverberant ratio, the performance of both algorithms is equivalent.

In spite of the stronger distortion of the signal, the strong improvement obtained in SNR gains and noise reduction is beneficial for automatic speech recognition. It can be hypothesized that the improvement in automatic speech recognition performance reflects the monaural algorithm's ability to reduce 'overlap masking' by cancelling out the energy in the signal which corresponds to the spreading by reverberation of previous phonemes.

## 5. Discussion

The main drawback of the algorithm is the presence of a residual musical noise. It can be efficiently reduced by the use of a spectral floor, to the detriment of the SNR gain and noise reduction performance. For automatic speech recognition application, the signal is intended for machine use only. Subjective quality of the speech signal is not relevant in this case. However, for hearing aid applications, the signal is presented to human ears after processing. A subjective test campaign would be the only way to confirm whether benefits can be brought about by the treatment whilst maintaining a tolerable subjective quality.

The model of late reverberation on which the algorithm is based is quite simplified. After the first echoes, the late reverberation part of real RIRs such as RIR1 to RIR6 used here, do exhibit an exponential decay behavior. However, the decay rate can vary along frequency. Moreover, the noise-like signal modulated by the exponential decay curve is only approximately white.

Direction for future work on the algorithm could be to improve this modeling. Differences in reverberation times along frequency can be readily integrated into the model, by estimating the reverberation time in sub-bands rather than globally. However, since the value of $T_r$ used in the algorithm appeared not to be critical, it is probable that the improvement gained would be small if compared with the increase in computational complexity.

An other limit of the model that could be addressed is the hypothesis that the signal modulated by a decaying exponential in the RIR is a white noise. This is a first approximation that does not account for the differences between RIR3, RIR5 and RIR6 for instance. The inversion of the minimum phase part of the RIR results in a whitening of the RIR [8]. To keep the algorithm blind, such a deconvolution could be realized as a first stage processing by cepstral deconvolution, prior to the monaural algorithm.

## 6. Conclusion

A novel algorithm for suppression of late reverberation from speech signals has been presented. It is based on amplitude spectral subtraction. Its novelty lies in the use of a model of the exponential decay of late reverberation. This model makes it possible to predict the PSD of reverberation, which can be then subtracted from the total PSD of the reverberated signal, yielding an estimate of the direct signal.

The system designed uses only one sensor. It can be added as a front end to other types of algorithms, such as automatic speech recognizer or cocktail party processors. Since the parameter of the model ($\Delta$) can be estimated on-line, the algorithm can automatically adapt to different rooms and acoustical situations. Moreover, $\Delta$ is related to $T_r$, the reverberation time, which is a characteristic of the room. Therefore, changes in the acoustic channel within the same room do not affect dramatically the performance of the algorithm. For applications such as distant sound pick-up automatic speech recognition, this means in practice that the user can move about the room while dictating to the speech recognizer, without it impairing the performance of the dereverberation

Table I. Bloom algorithm. NR- and G SNR-values are given in dB.

| RIR | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $T_r$ in s | 1.01 | 1.7 | 0.55 | 0.34 | 0.55 | 0.55 |
| NR | 3.9 | 4.4 | 4 | 3.6 | 4.4 | 4.8 |
| G SNR | 1.3 | 1.2 | 1 | 1.2 | 1.5 | 1.7 |
| CD | 0.05 | 0.06 | 0.05 | 0.05 | 0.03 | 0.04 |

Table II. Performance without on-line estimation of $T_r$. NR- and G SNR-values are given in dB.

| RIR | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $T_r$ in s | 1.01 | 1.7 | 0.55 | 0.34 | 0.55 | 0.55 |
| NR | 16.8 | 13.5 | 13.3 | 8.4 | 13.4 | 11.8 |
| G SNR | 0.8 | 0.6 | 0.2 | 0.3 | 1 | 0.7 |
| CD | 0.09 | 0.12 | 0.09 | 0.05 | 0.08 | 0.1 |

Table III. Performance with on-line estimation of $T_r$. NR- and G SNR-values are given in dB.

| RIR | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $T_r$ in s | 1.01 | 1.7 | 0.55 | 0.34 | 0.55 | 0.55 |
| estimated $T_r$ | 0.88 | 1.1 | 0.64 | 0.46 | 0.54 | 0.70 |
| NR | 16 | 10.3 | 14.5 | 10.5 | 13 | 13.2 |
| G SNR | 0.7 | 0.3 | 0.3 | 0.3 | 1 | 0.9 |
| CD | 0.08 | 0.1 | 0.09 | 0.06 | 0.08 | 0.1 |

Table IV. Speech recognition scores, Monaural and Bloom algorithms. "None" means . . .

| RIR | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| None | 41% | 25 % | 49% | 58% | 59% | 49% |
| Monaural | 78% | 65% | 75% | 76% | 73% | 77% |
| Bloom | 48% | 32% | 62% | 76% | 78% | 63% |

front end. It also means that the dereverberation will be the same whatever the position of the speaker. Hence in a cocktail party situation, the speech from people at different locations in the room will be equally dereverberated.

The algorithm achieves a strong reduction of the reverberant energy. It results in significant improvements in speech recognition scores, leading for all the RIR considered to recognition scores from 65% to 78%. For RIRs with long reverberation time and poor direct to reverberant ratio, significant improvements in speech recognition scores are achieved compared to a classical reference algorithm.

**Acknowledgement**

The authors whish to thank André Gilloire, in the CNET, Lannion, for the acquisition of the RIRs used in this work, as well as to thank the reviewers of this paper for their thorough review and their many helpful comments.

**References**

[1] Q. Lin et al.: Robust distant-talking speech recognition. ICASSP, 1996. 21–24.
[2] P. Castellano, S. Sridharan, D. Cole: Speaker recognition in reverberant enclosures. ICASSP, 1996. 117–120.
[3] A. K. Nabelek: Acoustical factors affecting hearing aid performance. Ed. G.A. Studebaker and I. Hochberg, 1993.
[4] A. V. Oppenheim, R. W. Schafer: Digital signal processing. Prentice Hall, 1975.
[5] S. T. Neely, J. B. Allen: Invertibility of room impulse response. J. Acoust. Soc. Am. **6** (1979) 165–169.
[6] H. Wang, F. Itakura: Dereverberation of speech signals based on sub-band envelope estimation. IEICE Transactions, 1991. 3576–3583.
[7] M. Miyoshi, Y. Kaneda: Inverse filtering of room impulse response. Proc. IEEE ASSP **36** (1988) 145–152.
[8] C. Marro: Traitements de déréverbération et de débruitage pour le signal de parole dans des contextes de communication interactive. Dissertation. Université de Rennes I, 1996.
[9] D. Cole, M. Moody, S. Sridharan: Position-independent enhancement of reverberant speech. J. Audio Eng. Soc. **45** (1997) 142–7.
[10] T. Langhans, H. W. Strube: Speech enhancement by non linear multiband envelope filtering. Int. Conf. on Acoust. Speech and Sig. Proc., 1982. 156–159.
[11] J. Mourjopoulos, J. Hammond: Modelling and enhancement of reverberant speech using an envelope convolution method. Int. Conf. on Acoust. Speech and Sig. Proc., 1983. 1144–1147.
[12] H. Hirsch, H. Finster: The reduction of reverberation to improve automatic speech recognition in rooms. 7th FASE Symposium, Edinburgh, Proceedings of Speech, 1988. 913–919.
[13] P. Clarkson, S. Bahgat: Envelope expansion methods for speech enhancement. J. Acoust. Soc. Am. **89** (1991) 1378–1382.
[14] P. Stringer, A. Tew: Binaural envelope expansion for speech dereverberation. Proceeding of the Institute of Acoustics, 1992. 57–64.
[15] W. Soede, F. A. Bilsen, A. J. Berkhout, J. Verschuure: Directional hearing aid based on array technology. Scandinavian Audiology **22 sup. 38** (1993) 20–27.
[16] W. Soede, A. J. Berkhout, F. A. Bilsen: Development of a directional hearing instrument based on array technology. J. Acoust. Soc. Am. **94** (1993) 785–796.
[17] M. Bodden: A concept for a cocktail party processor. Int. Conf. on Spoken Language Processing, 1990. 285–289.
[18] M. Bodden: Modeling human sound source localization and the cocktail-party-effect. Acta Acustica **1** (1993) 43–55.
[19] B. Kollmeier, J. Peissig, V. Hohmann: Binaural noise reduction hearing aid scheme with real-time processing in the frequency domain. Scandinavian Audiology **22** (1993) 28–38.
[20] B. Kollmeier, J. Peissig, V. Hohmann: Real-time multiband dynamic compression and noise reduction for binaural hearing aids. Journal of Rehabilitation Research and Development **30** (1993) 82–94.
[21] T. Wittkop, S. Albani, V. Hohmann, J. Piessig, W. S. Woods, B. Kollmeier: Speech processing for hearing aids: noise reduction motivated by models of binaural interaction. Acustica-Acta acustica **83** (1997) 684–699.
[22] A. Shamsoddini, P. Denbigh: A sound separation system for the hearing-impaired. Proceedings of the Institute of Acoustics, 1996. 19–26.
[23] Y. Cao, S. Sridharan, M. Moody: Speech seeking microphone array with multi stage processing. Eurospeech 95, 1995. 1991–1994.
[24] J. Allen, D. Berkley, J. Blauer: Multimicrophone signal-processing technique to remove room reverberation from speech signals. J. Acoust. Soc. Am. **62** (1977) 912–915.
[25] P. Bloom: Evaluation of a dereverberation process by normal and impaired listeners. Int. Conf. on Acoust. Speech and Sig. Proc., 1980. 500–503.
[26] P. Bloom, J. Cain: Evaluation of two input speech dereverberation techniques. Int. Conf. on Acoust. Speech and Sig. Proc., 1982. 164–167.
[27] R. Zelinski: A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. Int. Conf. on Acoust. Speech and Sig. Proc., 1988. 2578–2581.
[28] A. W. K.U. Simmer, S. Fischer: Suppression of coherent and incoherent noise using a microphone array. Annales des Télécommunications **49** (1994) 439–446.
[29] S. Fischer, K. U. Simmer: Beamforming microphone arrays for speech acquisition in noisy environments. Speech Communication **20** (1996) 215–27.
[30] C. Marro, Y. Mahieux, K. Simmer: Performance of adaptive dereverberation techniques using directivity controlled arrays. EUSIPCO, 1996.
[31] A. K. Nabelek et al.: Reverberant overlap- and self-masking in consonant identification. J. Acoust. Soc. Am. **86** (1989) 1259–1265.
[32] R. Bolt, A. MacDonald: Theory of speech masking by reverberation. J. Acoust. Soc. Am. **21** (1949) 577–580.
[33] J. Polack: La transmission de l'énergie sonore dans les salles. Dissertation. Université du Maine, 1988.
[34] J. Moorer: About this reverberation business. Computer Music Journal **3** (1979) 13–18.
[35] K. Lebart: Speech dereverberation applied to automatic speech recognition and hearing aids. Dissertation. Joint PhD, Université de Rennes I, University of Sussex, 1999.
[36] J. Allen, L. Rabiner: A unified approach to short-time fourier analysis and synthesis. Proceedings of the IEEE **65** (1977).
[37] B. Ayad: Systèmes combinés d'annulation d'écho acoustique et de réduction de bruit pour les terminaux mains-libres. Dissertation. Université de Rennes I, 1997.
[38] O. Cappé: Techniques de réduction de bruit pour la restauration d'enregistrements musicaux. Dissertation. Telecom Paris, 1993.
[39] Y. Ephraim, D. Malah: Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Transactions on A.S.S.P. **32** (1984) 1109–1121.
[40] S. Boll: Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on A.S.S.P. **27** (1979) 113–120.
[41] J. M. M. Berouti, R. Schwartz: Enhancement of speech corrupted by acoustic noise. ICASSP, 1979. 208–211.
[42] M. Schroeder: New method for measuring reverberation time. J.Acoust.Soc.Am. **37** (1965) 409–412.
[43] R.F.Kubichek: Standards and technology issues in objective voice quality assessment. Digital Signal Processing **1** (1991) 38–44.
[44] P. Bloom: Evaluation of a dereverberation process by normal and impaired listeners. ICASSP, 1980. 500–503.