REVERBERANT SOUND FIELD ANALYSIS USING A MICROPHONE ARRAY

Wolfgang Täger, Yannick Mahieux

CNET LAA/TSS/CMC
Avenue Pierre Marzin, F 22307 Lannion, France
tager@lannion.cnet.fr, mahieux@lannion.cnet.fr

ABSTRACT

The use of microphone arrays for sound pickup in reverberant environments has been proposed by many authors. The observation on the M microphones can be decomposed into a spatially coherent and an incoherent part. The first one is due to perfect (plane or spherical) sound waves caused by the direct path and specular reflections, whereas the latter is caused by diffusion, diffraction, non-perfect reflections, electrical and quantization noise. In this paper we firstly present a deflation method to detect and localize spatially coherent waves from the measured impulse responses. In a second step the filters which model the source directivity and the reflecting materials are estimated. The model takes into account nearfield delay, range attenuation, microphone and source directivity as well as non trivial reflections.

1. INTRODUCTION

The perceptual quality of a room is closely related to its reflections [4,5]. We present a method which combines the results of room simulation algorithms [1-3] with those of array processing in order to detect, localize, and analyse these reflections. The results are useful for acoustical room analysis as well as for the design of array processing algorithms.

2. NEARFIELD MODEL

The point-to-point room impulse response depends on the emitter and receiver positions and directivities, the room geometry and the properties of the reflections. Specular reflections can be modeled by "image sources" [1]. **Note that in the following "source" is used for real or image sources.** Let c be the sound velocity and $\mathbf{x_p}$ and $\mathbf{x_m}$ be the position vectors of the source p and the microphone m respectively. The corresponding path is affected by

1. Distance $\Rightarrow$ delay $\tau_{p,m} = \frac{\|\mathbf{x_p}-\mathbf{x_m}\|}{c}$
2. Distance $\Rightarrow$ attenuation $\frac{1}{\|\mathbf{x_p}-\mathbf{x_m}\|}$
3. Angle of arrival $\Rightarrow$ microphone directivity u_{x_p,x_m}
4. Emitter directivity $\Rightarrow$ direction dependent impulse response of the loudspeaker source
5. Reflection (wall,window ...) $\Rightarrow$ reflection filter

Let us suppose that the microphone positions, orientations and directivity patterns are known. We further assume that the directivity patterns of the microphones do not depend on frequency. This is ideally the case for omnidirectional or cardioid microphones. Obviously the first three alterations depend only on the relative position of the source. They induce a delay and a frequency independent gain, but do not affect the filter shape. They can be predicted by a suitable propagation model. The last two alterations are difficult to predict. The basic assumption in this paper is that they can be modelled by a single filter for each image source position, i.e. that they are independant of the microphone.

These assumptions lead to the following model : Each of the M point-to-microphone room impulse responses $h_m(n)$ is modelled by $\tilde{h}_m(n)$ which can be decomposed into a sum of filters $\tilde{h}_{p,m}(n)$ representing each one path. Each "path filter" $\tilde{h}_{p,m}(n)$ can be written as the convolution of two filters which are called "propagation filter" and "shape filter". We get :

The model and model error equation

$$h_m(n) = \tilde{h}_m(n) + e_m(n) \tag{1}$$

The multipath propagation equation

$$\tilde{h}_m(n) = \sum_{p=1}^{P} \tilde{h}_{p,m}(n) \tag{2}$$

The shape and propagation filter decomposition

$$\tilde{h}_{p,m}(n) = s_p(n) * g(\mathbf{x_p}, \mathbf{x_m}, n) \tag{3}$$

The propagation filter

$$g(\mathbf{x_p}, \mathbf{x_m}, n) = \alpha_{x_p,x_m}\,\delta(n - \tau_{p,m}) \tag{4}$$

with the frequency independent amplitude factor

$$\alpha_{x_p,x_m} = \frac{u_{x_p,x_m}}{\|\mathbf{x_p} - \mathbf{x_m}\|} \tag{5}$$

Finally we get the following nearfield model

$$\tilde{h}_m(n) = \sum_{p=1}^{P} \alpha_{x_p,x_m}\, s_p(n - \tau_{p,m}) \tag{6}$$

Let us examine the case that the impulse responses $h_m(n)$ have been measured. The problem can be stated as follows : Knowing these impulse responses, what can we know about the number of sources P, their positions $\mathbf{x_p}$ and the filters $s_p(n)$ who model the source directivity and the spectral shaping of the reflections ?

3. DETECTION AND LOCALIZATION

If we knew the position $\mathbf{x_p}$ of a given source p, then according to equation 3, a straightforward method to obtain the shape filter common to all microphones would be to pass each impulse response through the corresponding inverse propagation filters and to calculate the mean :

$$\frac{1}{M}\sum_m \frac{1}{\alpha_{x_p,x_m}} h_m(n+\tau_{p,m}) \tag{7}$$

This is a typical delay-weight-sum beamforming equation with the impulse response as the signal of interest. But this method would emphasize the microphones with the smallest signal levels. The results would be extremely unsatisfying if a source was close to a null direction of a single microphone. The best weighting in a least squares sense is proportional to the amplitude term α_{x_p,x_m}. This can be seen as the spatially matched filter. We obtain for n=0..K (d stands for deflation) :

$$s_{d,p}(n) = \frac{\sum_m \alpha_{x_p,x_m} h_m(n+\tau_{p,m})}{\sum_m \alpha^2_{x_p,x_m}} \tag{8}$$

In our application a shape filter length of $1.25ms$ is satisfying, corresponding to $N = K+1 = 20$ for a sampling frequency $F_s = 16kHz$. A sound wave propagates $42cm$ in this time. The array we have used is more than 3 times longer. So even if several reflections arrive simultaneously at one microphone, they do not completely overlap on all microphones unless the image sources are very close to each other in both distance and angle. To evaluate whether a point $\mathbf{x_e}$ is a source point or not we define :

The total energy :

$$T_e = \sum_{n=0}^{K}\sum_{m=1}^{M}(h_m(n+\tau_{e,m}))^2 \tag{9}$$

The residual energy :

$$E_e = \sum_{n=0}^{K}\sum_{m=1}^{M}(\alpha_{x_p,x_m} s_{d,e}(n) - h_m(n+\tau_{e,m}))^2 \tag{10}$$

The energy reduction ratio :

$$R_e = \frac{T_e}{E_e} \geq 1 \tag{11}$$

R_e indicates how well the model is able to describe the observation assuming that the source is located in point $\mathbf{x_e}$. A local maximum with a satisfying energy reduction of at least R_{min} is used as detection criterion. In our application the signal of interest is the shape s_p. We want to distinguish the signal + noise case $\mathbf{x_e} = \mathbf{x_p}$ from the noise only case $\mathbf{x_e} = \mathbf{x_b}$. We suppose that $\mathbf{x_b}$ is far away from $\mathbf{x_p}$ and that the noise $b_m(n)$ is spatially white, zero mean and non correlated with the signal. In a first time we suppose further that we have only one source $p = 1$ placed in $\mathbf{x_s} = \mathbf{x_{p=1}}$. In the sequel we derive the expectation value of R_e for $\mathbf{x_e} = \mathbf{x_s}$ and for $\mathbf{x_e} = \mathbf{x_b}$. The standard deviation of R_e for the noise case is estimated in section 5 by simulations.
For $\mathbf{x_e} = \mathbf{x_s}$ we obtain (the index p=1 is dropped) :

$$h_m(n+\tau_m) = \alpha_{x_s,x_m} s(n) + b_m(n) \tag{12}$$

$$s_d(n) = s(n) + \frac{\sum_m \alpha_{x_s,x_m} b_m(n)}{\sum_m \alpha^2_{x_s,x_m}} \tag{13}$$

Let P_s be the signal power, $\sigma^2_{b,m}$ be the noise power on microphone m and $P_{s,r} = g_{x_s} P_s$ the received signal power with

$$g_x = \sum_m \alpha^2_{x,x_m} \tag{14}$$

In the signal + noise case $\mathbf{x_e} = \mathbf{x_s}$, we obtain

$$E[T_e] = NP_{s,r} + N\sum_m \sigma^2_{b,m} \tag{15}$$

The noise only case can be derived by setting $P_{s,r} = 0$

$$E[T_e] = N\sum_m \sigma^2_{b,m} \tag{16}$$

In order to get a common formula for the transition, we introduce the steered energy $G_{x_e}P_s$ with $G_{x_e} = g_{x_s}$ for $\mathbf{x_e} = \mathbf{x_s}$ and $G_{x_e} = 0$ in the noise only case :

$$E[T_e] = G_{x_e}NP_s + N\sum_m \sigma^2_{b,m} \tag{17}$$

In both cases the residual energy is :

$$E[E_e] = N\sum_m (1 - \frac{\alpha^2_{x_e,x_m}}{g_{x_e}})\sigma^2_{b,m} \tag{18}$$

The transition between these cases depends on the spectral density of the signal since the beam patterns are frequency dependent. An analysis of this transition is out of the scope of this paper, we simply use the factor G_{x_e} which decreases from g_{x_e} to 0. Approximating the mean of the energy reduction ratio by the ratio of the means, we obtain :

$$E[R_e] \simeq \frac{G_{x_e}P_s + \sum_m \sigma^2_{b,m}}{\sum_m (1 - \frac{\alpha^2_{x_e,x_m}}{g_{x_e}})\sigma^2_{b,m}} \tag{19}$$

In the farfield all α_{x_e,x_m} are equal to $\sqrt{\frac{g_{x_e}}{M}}$. If the signal power is zero then the mean value of R_e depends only on the number of microphones M :

$$E[R_e] \simeq E_b = \frac{M}{M-1} \tag{20}$$

The same result E_b is obtained in the nearfield if the noise power is identical for all microphones. When a unique source is present with a mean SNR

$$SNR = \frac{P_{s,r}}{\sum_m \sigma^2_{b,m}} \tag{21}$$

then the expectation of R_e can be underestimated by

$$E[R_e] > E_s = \frac{P_{s,r} + \sum_m \sigma_{b,m}^2}{\sum_m \sigma_{b,m}^2} = SNR + 1 \qquad (22)$$

Let σ_{E_b} be the standard deviation of E_b. We define :

$$Q = \frac{E_s - E_b}{\sigma_{E_b}} = \frac{SNR - \frac{1}{M-1}}{\sigma_{E_b}} \qquad (23)$$

E.g. if we have M=4 microphones and a mean SNR of 0dB then the signal will yield $E_s = 2$ and the noise $E_b = 1.33$. For N=20 points, we found by simulation a standard deviation of $\sigma_{E_b} = 0.13$ so that the signal peak is $Q = \frac{E_s - E_b}{\sigma_{E_b}} = 5$ standard deviations higher than the noise mean. The probabilty of false detection is estimated lower than 10^{-5}.
How can we handle multiple reflections arriving simultaneously ? Can we expect to detect them ? If they are not too close, each reflection behaves somewhat like a noise for the detection of the other reflections. The worst case to detect at least one of P_0 sources is that all signals have the same received power. If the noise power is weak compared to the power of all signals together then the probability of detection depends on

$$Q = \frac{\frac{1}{P_0 - 1} - \frac{1}{M-1}}{\sigma_{E_b}} \qquad (24)$$

Once the first reflection has been detected, we also want to detect the other - possibly less powerful - reflections. Two methods are proposed in this paper. The first one is a straightforward extension of the unique source solution by a deflation method. Each time a source has been detected, the corresponding model is substracted

$$e_m(n) = h_m(n) - \tilde{h}_m(n) \qquad (25)$$

Then we reevaluate the region where the source has been found to avoid overlooking less powerful sources taking the residual e instead of h. Unfortunately the deflation of non orthogonal signals is known to modify them. To avoid this, we can use the localisation results and plug them into another algorithm which estimates again the shape filters by a more appropriate method. This can be done each time a source has been detected or after the complete detection. The algorithm is presented below.

4. SHAPE FILTER COMPUTATION

Once the number of sources P and their positions $\mathbf{x_p}$ have been estimated, the last unknowns are the shape filters, noted $s_{ls,p}$ (ls for least squares). They can be chosen in such a way that the model impulse responses are as close as possible to the observed ones in a least squares sense. We minimize the cost function :

$$J = \sum_{n=-\infty}^{\infty} \sum_m (h_m(n) - \tilde{h}_{ls,m}(n))^2 \qquad (26)$$

with

$$\tilde{h}_{ls,m}(n) = \sum_p \alpha_{x_p,x_m} s_{ls,p}(n - \tau_{p,m}) \qquad (27)$$

The non integer delay can be expressed using the interpolation function i(t) :

$$s_{ls,p}(n - \tau_{p,m}) = \sum_{k=0}^{K} s_{ls,p}(k) i(n - \tau_{p,m} - k) \qquad (28)$$

with eg.

$$i(t) = \frac{sin(\pi t)}{\pi t} = sinc(t) \qquad (29)$$

Inserting equations 27 and 28 into 26 we get :

$$J = \sum_{m,n,p,k} (h_m(n) - \alpha_{x_p,x_m} s_{ls,p}(k) i(n - \tau_{p,m} - k))^2 \qquad (30)$$

This is a quadratic form of the unknowns $s_{ls,p}(k)$. We can put all P(K+1) unknowns into a vector $\mathbf{s_{ls}}$ and write

$$J = \frac{1}{2} \mathbf{s_{ls}^t A s_{ls}} - \mathbf{b}^t \mathbf{s_{ls}} + c \qquad (31)$$

with

$$\mathbf{s_{ls}} = (s_{ls,1}(0), ..., s_{ls,1}(K), ..., s_{ls,P}(K))^t \qquad (32)$$

$$\mathbf{A} = \begin{pmatrix} a_{1,0,1,0} & \cdots & a_{1,K,1,0} & \cdots & a_{P,K,1,0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & a_{p,k,p_0,k_0} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{1,0,P,K} & \cdots & a_{1,K,P,K} & \cdots & a_{P,K,P,K} \end{pmatrix} \qquad (33)$$

$$a_{p,k,p_0,k_0} = \sum_m \alpha_{x_p,x_m} \alpha_{x_{p0},x_m} i(\tau_{p,m} + k - \tau_{p_0,m} - k_0) \qquad (34)$$

$$\mathbf{b} = (b_1(0), ..., b_1(K), ..., b_P(K))^t \qquad (35)$$

$$b_p(k) = \sum_m \alpha_{x_p,x_m} h_m(k + \tau_{p,m}) \qquad (36)$$

$\mathbf{A}$ is a (P(K+1),P(K+1)) real, symmetric, block-Toeplitz semipositive matrix. Excluding the singular case, the unique minimum of the cost function J is given by :

$$\mathbf{s_{ls}} = \mathbf{A^{-1}} \mathbf{b} \qquad (37)$$

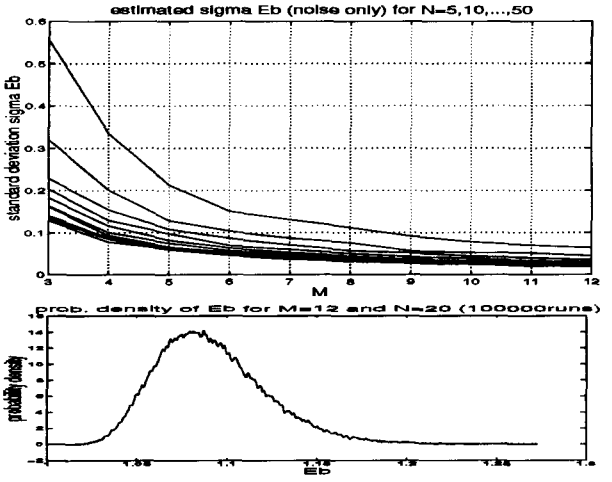
$\mathbf{A}$ is the mixing matrix and therefore $\mathbf{A^{-1}}$ is a separation matrix. If only one source is active and using sinc as interpolation function then the mixing matrix is $g_{x_p,x_m} I$. This proves that equation 8 is optimal in the least squares sense if only one source is active :

$$\mathbf{s_{ls}} = \mathbf{A^{-1}} \mathbf{b} = \mathbf{s_d} \qquad (38)$$

To calculate the reflection filters, we need to know the direction dependent impulse response of the loudspeaker source to perform the deconvolution. Only the angles of arrival have been estimated. Only a geometrical room model can give the corresponding send out angles. Therefore no results concerning this separation algorithm are given here. Note that this algorithm can also be applied to separate speech signals instead of impulse responses, or that we could use the cost function to design an adaptive localization using the results of the previous method as the starting point.

5. SIMULATION RESULTS

The standard deviation σ_{E_b} depends on M and N. We simulated spatially white noise sequences of length N on the M microphones, calculated R_e and estimated the mean and the standard deviation. The mean is very close to the predicted value given by equation 20. The standard deviation and a typical distribution are given below

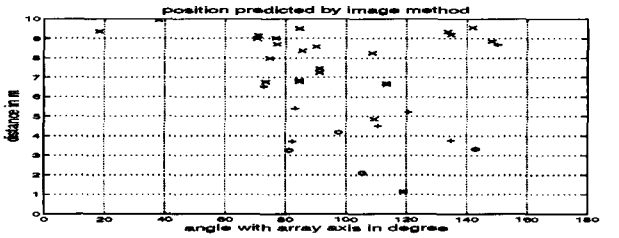


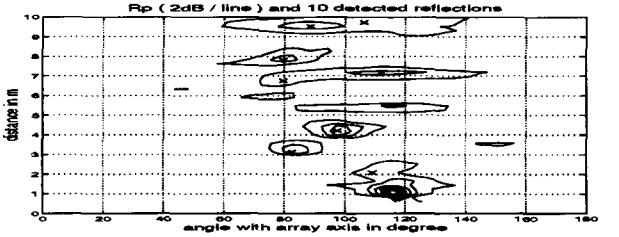
6. ROOM ANALYSIS RESULTS

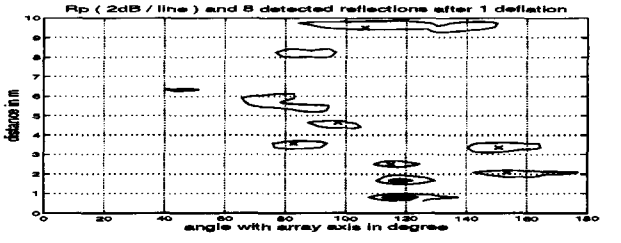
The room impulse responses have been measured for a 12-microphone logarithmic array with a total length of 1.44m. The small room is well damped, so that the reverberation time is only 180ms. Note that a small room is unfavorable in our case since reflections overlap more.

The comparison (see right column) of the image source model result (* for direct path, o for first order, + for second order, x for third order) with the first and second deflation of the energy reduction evaluation shows that (image method values between parentheses)

1. all first-order reflections have been found, even the floor reflection (2.10m, 105°) which must travel through the table on which the array was placed
2. the reflection $(3.27m, 81°)$ is separated from two reflections which are close in distance $(3.33m, 143°)$ and angle $(3.71m, 82°)$ respectively
3. for strong reflections the supposed shape filter length $N = 20$ may be too short so that they are detected again (direct path $1.15m, 119°$ and the strongest reflection at $4.18m, 98°$)
4. some objects in the room (table, chairs, second artificial head) caused non specular reflections, eg. at 2.10m, 152°. Note that the corresponding maxima are spread over larger angles







7. CONCLUSION

Two new algorithms have been developped to detect, localize and analyze reflections in a reverberant environment using a microphone array. The results can be used to choose an appropriate array processing method and to gain insight in the reflection characteristics of the materials. Measurements in a real room showed that even non-specular or multiple reflections can be detected and that reflections which arrive nearly at the same time can be separated. The number of detected reflections can be higher than the number of microphones. Even with few microphones and low SNR the probability of false detection is very low.

References

[1] Allen, Berkley , Image method for efficiently simulating small room acoustics , Journal of the Acoustical Society of America, April 79, pp. 943-950

[2] Emerit , Simulation binaurale de l'acoustique de salles de concert , Ph. D. Thesis, INPG France Sept. 95

[3] Christensen , The application of digital signal processing to large-scale simulation of room acoustics , Journal of the Audio Engineering Society, April 92, Vol. 40, Number 4, pp.260-276

[4] Kuttruff , Room Acoustics , Elsevier Applied Science 1991, 3rd edition

[5] Beranek , Music Acoustics and Architecture , John Wiley and Sons, 1962

[6] Tanaka, Kaneda , Performance of source direction estimation methods under reverberant conditions , Journal of the Acoustical Society of Japan Vol. 14, Number 4, 1993, pp. 291-292