Spectral and Spatial Multichannel Analysis/Synthesis of Interior Aircraft Sounds

Charles Verron*, Philippe-Aubert Gauthier**, Jennifer Langlois* and Catherine Guastavino**

Abstract—A method for spectral and spatial multichannel analysis/synthesis of interior aircraft sounds is presented. We propose two extensions of the classical sinusoids+noise model, adapted to multichannel stationary sounds. First, a spectral estimator is described, using average information across channels for spectral peak detection. Second, the residual modeling is extended to integrate two interchannel spatial cues (i.e., coherence and phase difference). This approach allows real-time synthesis and control of sounds spectral and spatial characteristics. It finds applications for multichannel aircraft sound reproduction, and more generally for musical and environmental sound synthesis. The ability of the model to reproduce multichannel aircraft sounds is assessed by a numerical simulation.

Index Terms—Aircraft sounds, sinusoidal modeling, spatial sound analysis/synthesis, spectral envelope, interchannel coherence, interchannel phase difference.

I. INTRODUCTION

Simulation of aircraft sounds has received a great interest in the last decades, mainly to provide test stimuli for studies of community annoyance to flyover noises. Synthesis methods have been proposed for reproducing such sounds by broadband, narrowband, and sinusoid components, including the time-varying aircraft position relative to the observer, directivity patterns, Doppler shift, atmospheric and ground effects [1], [2], [3]. Realistic synthesis of interior aircraft sounds has received less attention, despite its high potential for flexible sound generation in flight simulators. Sources of noise in aircraft cabins were reviewed in [4], [5]. The primary sources are the aircraft engines and the turbulent boundary layer noise. Promising approaches for simulating such sounds as combinations of sinusoidal and noisy components have been proposed in [6], [7], [8].

In this study we investigate a potential method for the analysis/synthesis of multichannel interior aircraft sounds. Compared to previous approaches, we aim at reproducing both spectral and spatial characteristics of the spatially-sampled or multichannel original sounds. For that purpose, we propose an extension to the sinusoids+noise analysis/synthesis model that takes into account two spatial cues: the interchannel coherence and the interchannel phase difference. Typically, the coherence is related to the “listener envelopment” or diffuseness of sound while the interchannel phase differences represent the spatial curvatures of the wave fronts that crosses the array. Our method allows analysis/synthesis of localized and diffuse sources, and more generally of any quasi-stationary multichannel signals (e.g., aircraft sounds, air conditioning noise, etc.) recorded with any multi-microphone technique (e.g., binaural manikin, Surround, high-order Ambisonics, arbitrary coincident and non-coincident microphone arrays and extended microphone arrays).

The analysis/synthesis model allows spectral and spatial parametric transformations of sound characteristics. Aircraft and vehicle comfort studies can greatly benefit from this possibility to modify the properties of specific sinusoidal or noisy components and investigate their impact on passengers’ perception. Binaural or multichannel stimuli are usually required for perceptual experiments in ecological listening conditions; such stimuli can be produced by the model since it applies to binaural and multichannel sounds. Other potential applications of the model (e.g., for controlling the spatial and spectral characteristics of multichannel sounds) will be presented in section V.

The paper is organized as follows: first we describe a set of descriptors for spectral and spatial multichannel characterization of aircraft sounds, then we present a multichannel sinusoids+noise analysis/synthesis scheme. Finally, the results of the model are evaluated by a numerical simulation and applications examples are described.

Throughout the paper the following notational convention is used: Scalars are denoted by normal letters (e.g., $x$ or $X$) and vectors and matrices using bold letters (e.g., $\mathbf{x}$ or $\mathbf{X}$). The $i$-th entry of a vector $\mathbf{x}$ is denoted by $[\mathbf{x}]_i$, and the $i, j$-th entry of a matrix $\mathbf{X}$ is denoted by $[\mathbf{X}]_{ij}$. Complex conjugate of the matrix $\mathbf{X}$ is represented by $\mathbf{X}^*$, and $\mathbf{X}^+$ denotes Hermitian transposition. In this document, a $Q$-channel recording is noted $\mathbf{x}(n)$ and the $q$-th channel is noted $[\mathbf{x}]_q(n) = x_q(n)$ where $n$ is the sample index.

II. CHARACTERIZATION OF AIRCRAFT SOUNDS

A. Sound database

An initial sound database has been provided by Bombardier and consists of binaural recordings made at different positions inside a CRJ900 Bombardier aircraft (see Figure 1). The data were recorded at 16 bits and 48 kHz.
The short-time Fourier transform (STFT) defined for each channel/synthesis scheme is designed for multichannel sounds and is not limited to binaural sounds. Though in the reminder of this paper, the term multichannel recordings can refer to binaural recordings (considered as multichannel with $Q = 2$) or more generally to recordings made with a microphone array.

To characterize aircraft recordings, we first compute the short-time Fourier transform (STFT) defined for each channel $x_q(n)$ by, $\forall m \in [0; M - 1], \forall k \in [0; N_a - 1]$:}

$$X_q(m, k) = \sum_{n=0}^{N_a-1} w_a(n) x_q(mM_a + n)e^{-j2\pi \frac{k}{N_a} n}$$

(1)

where $w_a$ is an analysis window of size $N_a$, $M_a$ is the analysis hopsize and $M$ is the total number of blocks.

To illustrate typical interior aircraft sounds, Figure 2 shows the magnitude of the STFT for a sound recorded in the aircraft at position 22SAD. It shows that interior aircraft sounds have quasi-stationary spectral properties. They contain sinusoidal components (spectral lines) and broadband background noise. Sinusoidal components are caused by rotating machinery and broadband noise is created by the turbulent boundary layer and the interior venting system. The presence of sinusoids with very close frequencies can be seen as slow amplitude modulation in the STFT. These are caused by asynchronous engine rotation and they are one of the salient signatures of interior aircraft sounds. In the following we assume that spectral properties of aircraft sounds are stationary. This assumption has been validated by a perceptual experiment [9] in which synthesized stationary sounds were indistinguishable from real aircraft recordings.

**B. Spectral and spatial multichannel descriptors**

Properties of interior aircraft sounds can be highlighted by computing a set of spectral and spatial descriptors. For each multichannel sound, we compute power spectra and two pair-wise interchannel cues: the interchannel coherence (IC) and the interchannel phase difference (IPD). Phase and amplitude differences between power spectra are related to the perceived position of localized sound sources, while interchannel coherence is related to the perceived source width or envelopment [10]. Such descriptors have been used in parametric multichannel audio coders [11], [12]. Due to the stochastic nature of aircraft sounds, a single short-time spectrum is not sufficient to estimate these quantities properly. However, assuming stationary spectral properties, the estimation can be done by averaging short-time spectra across time, as presented now.

1) **Spectral envelope:** The spectral envelope is estimated with the Welch’s method [13]. For each channel $q$, the STFT (calculated with an analysis hopsize $M_a = N_a/2$) is time-averaged to get the Welch power spectrum estimate:

$$\hat{X}_q(k) = \frac{1}{M} \sum_{m=0}^{M-1} | X_q(m, k) |^2$$

(2)

The spectral envelope is then defined as the square-root of the Welch power spectrum:

$$S_q(k) = \sqrt{\hat{X}_q(k)}$$

(3)

Note that spectral envelope and power spectrum are equivalent in decibel (dB), i.e., $20 \log(S_q) = 10 \log(\hat{X}_q)$. For...
that reason, both terms are used equivalently throughout this document.

2) Interchannel coherence: Several methods have been proposed to estimate the complex coherence [14] between two (or more) signals [15], [16], [17], [18], [19], [20]. In [15] the coherence is estimated by averaging short-time cross-spectra across time (similarly to the Welch’s method for power spectrum estimation). We use this method to estimate the complex coherence \( \gamma_{qp}(k) \) between each pair of channels \( q \) and \( p \):

\[
\gamma_{qp}(k) = \frac{1}{\sqrt{\sum_{m=0}^{M-1} |X_q(m,k)|^2 \sum_{m=0}^{M-1} |X_p(m,k)|^2}} \sum_{m=0}^{M-1} X_q(m,k) \overline{X_p(m,k)}
\]

(4)

Then the coherence function (also called magnitude squared coherence function) [21], [22] is given by:

\[
C_{qp}(k) = |\gamma_{qp}(k)|^2
\]

This function of frequency, with values between 0 and 1, gives a measure of correlation between each pair of channels, per frequency bin. Note that this coherence estimation method assumes that the analysis window size \( N_a \) is long compared to the time delay between channels, i.e., \( N_a \gg \frac{d_{mic}}{c f_s} \) where \( d_{mic} \) is the inter-microphone distance, \( c \) the speed of sound, and \( f_s \) the sampling frequency.

3) Interchannel phase difference: The interchannel phase difference is related to the curvature of the wave fronts crossing the array. It is computed between all pairs of channels by:

\[
P_{qp}(k) = \angle \gamma_{qp}(k)
\]

(5)

4) Matrix notation: For analysis/synthesis of multichannel signals in section III-C it will be useful to express the spectral envelope, interchannel coherence and phase difference in a compact matrix form. For that purpose, we define the matrix \( \hat{X}(k) \in \mathbb{C}^{Q \times M} \) whose \( q,m \)-th element is \( \hat{x}_{qm}(k) = X_q(m-1,k) \):

\[
\hat{X}(k) = \begin{bmatrix}
X_1(0,k) & \cdots & X_1(M-1,k) \\
\vdots & \ddots & \vdots \\
X_Q(0,k) & \cdots & X_Q(M-1,k)
\end{bmatrix}_{Q \text{ channels}}^{M \text{ time frames}}
\]

The spectral envelope vector \( \hat{S}(k) \in \mathbb{R}^{+Q \times 1} \) is then defined by:

\[
[\hat{S}]_{q}(k) = S_q(k) = \sqrt{1/M} |\hat{X}^* \hat{X}|_{qq}(k)
\]

(6)

Similarly we define the interchannel coherence matrix \( \hat{C} \in \mathbb{R}^{+Q \times Q} \) by:

\[
[\hat{C}]_{qp}(k) = |\hat{X}^* \hat{X}|_{qp}(k) / |\hat{X}^* \hat{X}|_{qq}(k)
\]

(7)

and the interchannel phase difference matrix \( \hat{P} \in \mathbb{R}^{Q \times Q} \) by:

\[
[\hat{P}]_{qp}(k) = \angle |\hat{X}^* \hat{X}|_{qp}(k)
\]

(8)

Remark: Definitions 6, 7 and 8 show that spectral envelopes, interchannel coherences and phase differences are entirely defined by \( \hat{X}^* \hat{X} \). This implies that two multichannel signals \( x \) and \( \hat{x} \) have the same spectral envelopes, interchannel coherences and phase differences if their STFT matrices satisfy:

\[
\hat{X}^* \hat{X} = \hat{X} \hat{X}^*
\]

(9)

This property will be exploited for analysis/synthesis of multichannel signals in section III-C.

C. Analysis window

Figure 3 shows the spectral envelopes, interchannel coherence and phase difference for two binaural aircraft sounds \( (N_a = 8192) \). The binaural recordings were made at row 04 and 22 in the aircraft cabin.

Figure 3. Spectral envelopes in [dB ref 1], interchannel coherence and phase difference for two binaural aircraft sounds \( (N_a = 8192) \). The binaural recordings were made at row 04 and 22 in the aircraft cabin.
choose $f_c = \frac{3.5}{N_a}$, resulting in a DPSW having nearly all its energy contained in its first lobe which is $K = 7$ points large in the $N_a$-point discrete Fourier transform of the window.

### III. Synthesis Model

Based on the stationary spectral and spatial characterization of aircraft sounds, we derive a multichannel sinusoids+noise analysis/synthesis model. The sinusoidal components are detected by means of a multichannel spectral estimator, and removed (by direct zeroing) from the original sound. Then the residual is modeled in terms of spectral envelopes, IC and IPD cues. The general process is illustrated on figure 4.

#### A. Sinusoids+noise model

In [24] an analysis/synthesis system based on a deterministic plus stochastic decomposition of a monophonic sound is presented. The deterministic part $d(t)$ is a sum of $I(t)$ sinusoids whose instantaneous amplitude $a_i(t)$ and frequency $f_i(t)$ vary slowly in time:

$$d(t) = \sum_{i=1}^{I(t)} a_i(t) \cos \left( \int_0^t 2\pi f_i(\tau) d\tau + \varphi_i \right)$$

where $\varphi_i$ is the sinusoid initial phase. The stochastic part $s(t)$ is modeled as:

$$s(t) = \int_0^t h(t, \tau)b(\tau)d\tau$$

where $b(t)$ is a white noise input and $h(t, \tau)$ the impulse response of a “time-varying” filter. This deterministic plus stochastic modeling (also called sinusoids+noise model) has been used extensively for analysis, transformation and synthesis of speech, musical and environmental sounds (see for example [25]). The typical steps of the analysis are:

- STFT: Compute the short-time Fourier transform.
- Peak picking: Detect prominent peaks in the STFT.
- Partial tracking: Connect peaks together to form partials with time-varying frequency and amplitude.
- Residual modeling: Remove partials from the original sound and estimate the residual spectral envelope.

Then the synthesis of the deterministic and stochastic components are performed:

- Partially are resynthesized with appropriate time-varying frequency, amplitude and phase.
- White noise is modulated in the frequency domain with the appropriate spectral envelope.

Modifications can be applied to the synthesis parameters, resulting in high-quality parametric transformations such as pitch-shifting, time-stretching, morphing, etc.

Note that in our case the model is simplified due to the stationarity hypothesis: the number of sinusoids, their frequency and the noise spectral envelope are constant in time. Our contribution is to extend the analysis and synthesis processes to the multichannel case. When dealing with multichannel sounds, extracting relevant information from all channels can improve the reliability of the analysis. This was used for binaural sounds in [26] to improve partial tracking. Here we propose a multichannel peak detection method that uses an average information across channels to improve the signal/noise ratio. Second, we present a multichannel approach for residual modeling, which includes IC and IPD cues. This method allows to analyze and resynthesize spatial characteristics of the noisy components.

#### B. Multichannel extraction of stationary sinusoidal components

1) Multichannel peak detection: Various estimation methods have been proposed in the literature for power spectrum estimation and sinusoidal detection (see for example [13], [27], [28], [29], [30], [31]). The modified periodogram first proposed by Welch [13] is particularly appropriate to detect sinusoidal peaks in aircraft sounds, due to their long-term quasy-stationary properties identified above. When averaging the signal STFT to compute the Welch spectral estimate (see Section II), the variance of noisy components decreases with the number of averaged blocks, while sinusoidal peaks remain constant. This reduces the signal/noise ratio and avoids false peak detection in noisy spectral components. A compromise must be established between the size of the analysis window (i.e., the frequency resolution) and the number of averaged blocks in the STFT (i.e., variance of noisy components) depending on the length of the available signals. Reproducing interior aircraft sounds requires a very good frequency resolution since sinusoidal components coming from the engines are sometimes separated by only a few Hertz. To capture these nearly-coincident sinusoids, we use a 2-second analysis window, assuming that long signals are available (e.g., 16 seconds). Results of this approach will be detailed in section IV.

In a multichannel aircraft sound, sinusoidal components are expected to be present in all channels, at the same frequency but with different amplitudes and phases. They could be identified by detecting the prominent peaks in the Welch spectral estimate $X_q(k)$ of each channel $q$. However,
the estimated peak frequencies may vary from one channel to another due to the stochastic perturbations. At the synthesis stage, these slight variations result in strong amplitude modulation distortion artifacts (beating components). Also, a per-channel peak detection is likely to miss the detection of sinusoids with low amplitude in one channel. To avoid these artifacts, we propose to perform the peak detection on a single spectral estimator, averaged across the Q channels:

$$\hat{X}(k) = \frac{1}{Q} \sum_{q=1}^{Q} \hat{X}_q(k) = \frac{1}{Q} \sqrt{\frac{1}{M} \text{Tr}(XX^*)}(k) \quad (10)$$

where $\text{Tr}(A)$ indicates the trace of the matrix $A$ which is the sum of the main diagonal elements. A sinusoid is detected at frequency $k_i$ if three conditions are satisfied:

- $X(k_i)$ is a local maximum, i.e., $\hat{X}(k_i - 1) < \hat{X}(k_i) < \hat{X}(k_i + 1)$
- $\forall k \in [k_i - 2K, k_i] \mid 10 \log(\hat{X}(k_i)) > 10 \log(\hat{X}(k)) + \delta$
- $\forall k \in [k_i - 1, k_i + 2K] \mid 10 \log(\hat{X}(k_i)) > 10 \log(\hat{X}(k)) + \delta$

where $\delta$ is the peak detection threshold, and $K$ denotes the size of the spectral lobe of the DPSW analysis window (see section II-C). This way, a peak is detected if two points in the neighborhood of a local maximum are under the given threshold. The size of the neighborhood was fixed to $4K$, i.e., four times the spectral width of the DPSW.

Using the average estimator (eq. 10) ensures that sinusoids are detected at the same frequency in all channels. Furthermore, since the sinusoidal components are coherent from one channel to another, contrary to the diffuse stochastic components, the signal to noise ratio (i.e., the sinusoid to noise ratio) is also improved by averaging across channels. The signal to noise ratio improvement can be determined by considering the variance of the estimator:

$$\text{Var} \left( \hat{X}(k) \right) = \text{Var} \left( \frac{1}{Q} \sum_{q=1}^{Q} \hat{X}_q(k) \right) = \frac{1}{Q^2} \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \text{Cov} \left( \hat{X}_q(k), \hat{X}_{q'}(k) \right)$$

We can assume that the $Q$ single-channel spectral estimators $\hat{X}_q(k)$ have similar variances, i.e., $\text{Var} \left( \hat{X}_q(k) \right) \approx \nu(k)$ for all $q$. If the $Q$ channels are completely correlated, which is the case for deterministic sound fields (i.e., containing only pure sinusoids) and sound fields coming from a single direction, then $\text{Cov} \left( \hat{X}_q(k), \hat{X}_{q'}(k) \right) \approx \nu(k)$ for all $(q, q')$. In that case averaging across channels has no effect on the variance, i.e., $\text{Var} \left( \hat{X}(k) \right) = \text{Var} \left( \hat{X}_q(k) \right)$. On the other hand, when the $Q$ channels are completely uncorrelated, which is the case for a diffuse stochastic sound field (i.e., a sound field made of many decorrelated plane waves) then $\text{Cov} \left( \hat{X}_q(k), \hat{X}_{q'}(k) \right) = 0$ if $q \neq q'$, so that the variance of the average estimator $\hat{X}$ is decreased by a factor $Q$:

$$\text{Var} \left( \hat{X}(k) \right) = \frac{1}{Q} \text{Var} \left( \hat{X}_q(k) \right)$$

As a result, in diffuse sound fields, the multichannel estimator greatly facilitates the peak detection compared to a single channel estimator. Another interpretation of this result can be stated as follows. For stochastic components, the variance of the Welch estimate is approximately inversely proportional to the number of blocks $M$ used for time-averaging [13]. When the $Q$ channels contain decorrelated signals, then averaging the estimator across the $Q$ channels is equivalent to increasing the number of blocks from $M$ to $QM$, hence reducing the variance by a factor $Q$.

Finally, note that the peak detection algorithm could be completed by perceptual criteria, e.g., to exclude some tones due to masking considerations. The method could be extended by a post-processing stage to determine the prominence from a perceptual point of view, extending the prominence ratio method [32] to the multichannel case. Future directions also include the use of other high-resolution techniques [14] to detect the spectral peaks. In particular, methods based on coherences rather than power spectral densities should be investigated as they may be advantageous for filtering out a significant part of the diffused noise.

2) Sinusoidal extraction and resynthesis: After peak detection, the detected sinusoid normalized frequency is $f_i = \frac{k_i}{N_a}$ (where $N_a$ is the size of the analysis window). The amplitude of the complex exponential is estimated in each channel $q$ by $a_{q,i} = S_q(k_i)$. Note that the multichannel average estimator $\hat{X}$ and $S_q$ are computed with the same STFT parameters. When computing the STFT, peaks of the complex exponentials are implicitly convolved with the spectrum of the analysis window $w_n$. For this reason, $w_n$ is normalized to have a unity sum (i.e., $\sum_{n=-\infty}^{+\infty} w_n(n) = 1$) so that its spectrum is equal to 1 at $f = 0$ Hz.

To determine the phase of each detected sinusoid, a fast Fourier transform (FFT) $F_q(k)$ is processed on the entire signal available for each channel $q$. The total number of available samples per channel is noted $N_{\text{tot}}$. The phase of each detected sinusoid is estimated by $\varphi_{q,i} = \angle F_q(k_i')$ where $k_i'$ is the nearest integer to $(f_i \cdot N_{\text{tot}})$. Finally, to remove the sinusoid from the original sounds, $F_q(k)$ is forced to 0 on $\left( k_i' + 1 \right) \frac{N_{\text{tot}}}{N_a}$ bins around $k_i'$, where $K$ is the spectral width of the DPSW analysis window (see section II-C). Then the inverse fast Fourier transform (IFFT) is processed to get the multichannel residual signal $r(n) \in \mathbb{R}^{Q \times 1}$ which contains the background noise, while the multichannel deterministic component is resynthesized by:

$$d(n) = 2R \begin{bmatrix} a_{11}e^{j\varphi_{11}} & \ldots & a_{1L}e^{j\varphi_{1L}} \\ \vdots & \ddots & \vdots \\ a_{Q1}e^{j\varphi_{Q1}} & \ldots & a_{QL}e^{j\varphi_{QL}} \end{bmatrix} \begin{bmatrix} e^{2\pi j f_1 n} \\ \vdots \\ e^{2\pi j f_L n} \end{bmatrix}$$

$$\text{(11)}$$

C. Multichannel residual modeling

The residual signal $r(n)$ is modeled as a multichannel stationary stochastic process $s(n)$. In the classical monophonic case, the synthesized stochastic process is computed...
to have the same spectral envelope as \( r(n) \). Here, this condition must be fulfilled for each channel. In addition, the synthesized multichannel signal should conserve the spatial characteristics of the residual, i.e., the channels of \( s(n) \) should have the same pair-wise IC and IPD as the channels of \( r(n) \). Given these conditions, the residual modeling is done in two stages: first, spectral and spatial descriptors are estimated from the residual with the Welch method as described in Section II. Second, the multichannel stochastic process is synthesized in the frequency domain, as a linear combination of \( Q \) synthetic filtered noises, satisfying the spectral and spatial constraints defined above.

1) Multichannel analysis: The \( q \)-th channel of the residual is noted \( r_q(n) \) and its STFT \( \hat{R}_q(m,k) \). We also define the matrix \( \mathbf{R} \in \mathbb{C}^{Q \times M} \) by \( |\hat{R}_{qm}(k)| = R_q(m,k) \). The \( Q \) residual power spectra \( |\mathbf{S}_q(k)| \) are computed from \( \mathbf{R} \) as defined in Section II, along with interchannel spatial cues \( |\mathbf{C}_{qp}(k)| \) (coherence) and \( |\mathbf{P}_{qp}(k)| \) (phase difference). The STFT analysis window \( w_q \) is normalized to have a unity power (i.e., \( \sum_{n=-\infty}^{\infty} w_q^2(n) = 1 \)) to compensate its global effect on power spectra magnitudes [13].

2) Multichannel synthesis: The synthesized stochastic components are constructed in the time-frequency domain, as proposed in [24], [33] for monophonic signals. Each channel \( q \) is defined as a synthetic STFT \( \hat{R}_q(m,k) \) and for each frame \( m \) the IFFT is computed by:

\[
\hat{r}_q^{(m)}(n) = \frac{1}{\sqrt{N_s}} \sum_{k=0}^{N_s-1} \hat{R}_q(m,k) e^{2j\pi \frac{k}{N_s}n} \quad \forall n \in [0; N_s - 1]
\]

where \( 1/\sqrt{N_s} \) is a normalization coefficient to conserve the energy of the stochastic process, \( N_s \) is the synthesis window size and \( M_s \) is the hopsize. The resulting short-time segments \( \hat{r}_q^{(m)} \) are overlapped and added to get the full time-domain synthesized channel \( s_q(n) \):

\[
s_q(n) = \sum_{m=0}^{M-1} w_q(n - mM_s) \hat{r}_q^{(m)}(n - mM_s)
\]

where \( M \) is the number of frames of the synthesized signal. For a wideband stochastic process, the overlap-add synthesis window should respect \( \sum_{n=-\infty}^{+\infty} w_q^2(n - mM_s) = 1 \) \( \forall n \) so that the synthesized signal has a constant variance [30]. We take the square root of the Hanning window with a synthesis hopsize \( M_s = \frac{N_s}{2} \).

The monophonic synthesis process described above must be completed in our case, such that each pair of synthesized channels satisfies the IC and IPD relations extracted at the analysis stage. Let \( \hat{\mathbf{R}}(k) \in \mathbb{C}^{Q \times M} \) be the matrix containing the synthetic short-time spectra, i.e., \( |\hat{R}_{qm}(k)| = R_q(m,k) \). In this matrix the number of columns \( M \) can be arbitrarily long (i.e., the synthesized signals are potentially infinitely long) but in practice the matrix is constructed frame by frame in real time. Given the definitions of Section II, to preserve the stationary spectral properties of the multichannel residual, along with its spatial properties, the synthetic matrix \( \hat{\mathbf{R}}(k) \) should satisfy eq. 9, which gives (when taking into account the length of the signals):

\[
\frac{1}{M} \hat{\mathbf{R}}^*(k) = \frac{1}{M} \mathbf{R}^*(k)
\]

A solution to this equation was given in [34] for the binaural case (\( Q = 2 \)) by:

\[
\hat{\mathbf{R}}(k) = \mathbf{A}(k) \begin{bmatrix} G_1(1,k) & \ldots & G_1(M,k) \\ G_2(1,k) & \ldots & G_2(M,k) \end{bmatrix}
\]

where

\[
\mathbf{A}(k) = \frac{M}{M} \begin{bmatrix} S_1(k) & 0 \\ S_2(k) \sqrt{C_{21}(k)} & S_4(k) \sqrt{1 - C_{21}(k)} \end{bmatrix}
\]

and \( G_1(m,k) \) and \( G_2(m,k) \) are the STFT of two Gaussian random sequences. To find an appropriate matrix \( \mathbf{A} \) in the generalized \( Q \)-channel case, we introduce the \( Q \times Q \) matrices \( \mathbf{U} \) and \( \mathbf{D} \) that diagonalize \( \mathbf{R}^*(k) \):

\[
\mathbf{R}^*(k) = \mathbf{U} \mathbf{D} \mathbf{U}^*(k)
\]

and compute \( \mathbf{A}(k) \) by:

\[
\mathbf{A}(k) = \frac{1}{\sqrt{M}} \mathbf{U}(k) \sqrt{\mathbf{D}(k)}
\]

where \( \mathbf{U} \) is a unitary matrix of eigenvectors, \( \mathbf{D} \) is a diagonal matrix of eigenvalues stored in increasing order, and \( \sqrt{\mathbf{D}} \) denotes elementwise square root of \( \mathbf{D} \). Note that the computation of \( \mathbf{A}(k) \) is realized only once, during the offline analysis stage.

At run-time, the matrix of synthesized short-time spectra \( \hat{\mathbf{R}}(k) \) is computed by:

\[
\hat{\mathbf{R}}(k) = \mathbf{A} \mathbf{G}(k)
\]

where \( \mathbf{G}(k) \in \mathbb{C}^{Q \times M} \) is a matrix whose each line \( G_q(m,k) \) is the STFT of a Gaussian noise sequence. In practice, \( \mathbf{G}(k) \) is synthesized efficiently in the frequency domain by computing its \((q,m)\)-th element as:

\[
|G|_{qm}(k) = G_{qm}(m,k) + j \sigma_{q3}(m,k)
\]

where \( j^2 = -1 \), and \( G_{qm}(m,k) \) and \( \sigma_{q3}(m,k) \) are independent real-valued Gaussian sequences with zero mean and standard deviation \( \sigma_{q}(k) \) and \( \sigma_{q3}(k) \) equal to \( \frac{1}{\sqrt{2}} \). Since all spectra are conjugate symmetric we only consider positive frequencies (i.e., \( k \in [0; \frac{N_s}{2}] \)). Also, for \( k = 0 \) or \( \frac{N_s}{2} \) (i.e., Zero and Nyquist frequency bins) \( \sigma_{q}(k) = 1 \) and \( \sigma_{q3}(k) = 0 \) since the time-domain synthesized signals are real-valued.

3) Validation: All coefficients of \( \mathbf{G}(k) \) are computed from different noise sequences, so when \( M \rightarrow \infty \) we have:

\[
\mathbf{G} \mathbf{G}^*(k) = (\sigma_{q}^2 + \sigma_{q3}^2) M \mathbf{I}_Q
\]

where \( \mathbf{I}_Q \in \mathbb{R}^{Q \times Q} \) is the identity matrix. It follows from eqs. 17, 18, 15 and 16 that:

\[
\hat{\mathbf{R}} \hat{\mathbf{R}}^*(k) = (\mathbf{A} \mathbf{G})(\mathbf{A} \mathbf{G})^*(k)
\]

\[
= \frac{M}{M} \mathbf{R} \mathbf{R}^*(k)
\]
which guarantees that the synthetic short-time spectra preserve the spectral and spatial cues of the analyzed multichannel residual \( r(n) \) (see eq. 13).

The synthesis parameters (i.e., spectral envelopes, coherence and phase difference) can also be modified to reproduce sound fields with different directivity/diffusion and spectral content. Such modifications are included in the computation of \( \mathbf{A}(k) \) and do not affect the real-time capabilities of the synthesis process. The proposed formulation is general, it allows to synthesize any multichannel stochastic signal whose channels satisfy spectral characteristics, along with pair-wise coherences and phase differences. This has applications for multichannel aircraft sound synthesis, and more generally for creating multichannel environmental and musical sounds. Applications examples are given and discussed in Section V.

4) Time-frequency resolution and computational cost: The choice of the residual analysis/synthesis window size \( N_s \) is important since it directly impacts the time/frequency resolution and the computational complexity of the synthesis. If we assume a hopsize of \( \frac{N_s}{f_s} \) samples and a sampling frequency \( f_s \), the corresponding time-resolution (latency) is \( \frac{N_s}{f_s} \) seconds, while the spectral resolution is \( \frac{f_s}{N_s} \) Hz. If we consider that an IFFT requires \( 2N_s \log_2(N_s) \) real multiplications per sample for each channel. An additional cost of \( Q^2 \) multiplications per sample (independent of \( N_s \)) is required for the multiplication by \( \mathbf{A}(k) \). Since the spectral properties of aircraft sounds are stationary, a long window is appropriate to guaranty a good frequency resolution. However, if the synthesis parameters are to be modified, e.g., to simulate different flight conditions, or a different recording position in the cabin, then a short window is desirable to ensure real-time transitions between the different synthesis conditions. Also, computational cost encourage one to use shorter window sizes. A reasonable time/frequency/computation tradeoff can be achieved by window sizes ranging from \( N_s = 128 \) to 8192 samples (i.e, 2.67 to 21.33 milliseconds with \( f_s = 48 \) kHz). The effects of the window size on the perceived synthesis quality were further investigated in [9]. It was found that for binaural aircraft signals, an appropriate window size is 1024 samples (with \( f_s = 48 \) kHz), leading to a time resolution of 10.66 milliseconds, a frequency resolution of 46 Hz (sufficient for broad-band coloration reconstruction) and a computational complexity of 80 multiplications per sample per channel (plus \( Q^2 \) multiplications per sample for the matrixing).

IV. EVALUATION OF THE MODEL

A numerical simulation was conducted to assess the capabilities of the model in the multichannel case (\( Q = 80 \)). Controlled numerical signals were used for the evaluation of the model to ensure precise and \textit{a priori} known types of sound field. We defined two simulation cases that are plausible in a real aircraft situation, namely a localized source and a diffuse sound field. The simulated microphone array is shown in Figure 5. This array configuration corresponds to a real microphone array studied and tested in [35]. We simulated the multichannel sound \( \mathbf{x}(n) \) received at the microphone array in the presence of spatialized deterministic and stochastic components. The deterministic component aimed at simulating the sinusoids emitted by the aircraft engines, while the stochastic component represented the noisy sources in the aircraft, such as aerodynamic noise and air-cooling systems. Two cases were investigated for the spatial distribution of the noisy component:

1) A localized sound source (a single plane wave of white noise).
2) A diffuse-like sound field (made of 30 plane waves of decorrelated white noise).

Both spatial distributions are plausible in a real aircraft cabin. Note that the diffuse-like sound field is not a diffuse sound field in the most strict sense for all the frequency range. Our goal here is not to estimate the incoming direction of the sinusoidal plane waves in the noisy sound field. We want to perform a multichannel analysis/synthesis of the signal \( \mathbf{x}(n) \) and assess:

- The ability of the average spectral estimator \( \hat{X} \) to detect sinusoids with very close frequency.
- The ability of the analysis/synthesis process to reconstruct the spectral and spatial properties of the original sound field \( \mathbf{x}(n) \), in terms of spectral envelopes, inter-channel coherence and phase differences.

We now present the theoretical construction of the multichannel stimuli \( \mathbf{x}(n) \) in detail and the results of the multichannel analysis/synthesis scheme.

A. Stimuli

We used a theoretical array of \( Q = 80 \) omnidirectional microphones, corresponding to the real array proposed in [35], [36], [37]. The array width and length are 0.993 m and 1.3475 m respectively, and the array height is 12.25 cm. The multichannel impulse response of the array \( \mathbf{h}_u(n) \) was calculated for \( U = 30 \) propagating directions, with azimuth \( \theta_u \) and elevation \( \alpha_u \). The \( Q \)-channel microphone array and the \( U \) plane wave directions are depicted on Figure 5. Ideal free field conditions were assumed. The \( q \)-th microphone is located by a vector \( e_q = [e_{1q}, e_{2q}, e_{3q}]^T \) (in meters) describing its position in rectangular coordinates. Assuming a sound speed \( c = 343 \) m/s, the time delay matrix \( \Delta \in \mathbb{R}^{Q \times U} \) for the plane waves impinging on the microphones is given by:

\[
[\Delta]_{qu} = \frac{1}{c} e_q \cdot \mathbf{n}_u \quad (19)
\]

where \( \mathbf{n}_u = [\cos(\theta_u) \cos(\alpha_u) \sin(\theta_u) \cos(\alpha_u) \sin(\alpha_u)]^T \) is a unitary vector aligned with the \( u \)-th propagation direction. The largest negative delay of \( \Delta \) is then identified to create an overall time offset that makes the smallest time delay equal to zero. Since these time delays are fractional sample delays, the evaluation of \( \mathbf{h}_u(n) \) is achieved using Lagrange interpolation [38] of order 31, thus creating 32-sample long
finite-impulse-response (FIR) filters. The remaining integer parts of the delay \((\Delta - 16)\) are processed as simple integer delay buffers. Lagrange interpolation is a conventional approximation method for fractional delays. It is also the easiest method to approximate a fractional delay. Hence, for the purpose of this evaluation this simple method is selected. However, although Lagrange interpolation filters are made maximally flat at a certain frequency (typically at 0 Hz), these FIR filters might alter the gain response of \(h_u(n)\) around the Nyquist frequency [38]. In this paper, the Lagrange interpolation order was selected to ensure a good compromise between 1) magnitude and phase responses for a wide frequency range from 0 Hz (where the response is maximally flat) and 2) low-order FIR filter for computation efficiency. In our case, the magnitude deviation of the FRFs starts to be important from 14 kHz. The worst magnitude deviation is in the range of -31 dB ref 1 at the Nyquist frequency. These limitations do not compromise the validity of the evaluation.

Two multichannel signals \(x^{(L)}(n)\) and \(x^{(D)}(n)\) (the superscripts \(L\) and \(D\) denote the localized (in terms of propagating direction) and diffuse cases respectively) were created as a sum of deterministic and stochastic sound sources. In both cases, the deterministic source was made of two complex exponential plane waves:

\[
p_i(n) = a e^{j2\pi\frac{f_i}{f_s}n}, \quad i = 1, 2
\]

with \(a = 0.01\), \(f_1 = 999\) Hz and \(f_2 = 1000.5\) Hz. We choose these frequencies so as to simulate the very slight dissimilarities frequently found in the speed of the two aircraft engines. To simulate the position of the engines, the incoming direction \((\theta, \alpha)\) of the sinusoidal plane waves were respectively equal to \((-30, 0)\) and \((30, 0)\) degrees. The resulting multichannel deterministic signal was computed as:

\[
d(n) = (p_1 \ast h_{\theta=-30} + p_2 \ast h_{\theta=30})(n) \tag{20}
\]

The localized noisy source was made of a plane wave in the horizontal plane with incoming direction \(\theta = 0\) degrees. The resulting signal at the microphones (deterministic plus stochastic contributions) was calculated by:

\[
x^{(L)}(n) = d(n) + (s_0 \ast h_{\theta=0})(n) \tag{21}
\]

where \(s_0(n)\) is a Gaussian white noise with variance \(\sigma^2 = 1\), filtered by a second-order ARMA filter to produce a pink noise more realistic for aircraft sounds (the filter coefficients for producing pink noise are given in [39]). The diffuse stochastic sound source was created as a sum of \(U = 30\) plane waves with different incoming directions. The resulting multichannel signal was calculated by:

\[
x^{(D)}(n) = d(n) + \sum_{u=1}^{U} (s_u \ast h_{\theta_u})(n) \tag{22}
\]

where the signals \(s_u(n)\) were decorrelated Gaussian white noises with variance \(\frac{\sigma^2}{U}\), filtered with the second-order ARMA filter. Note that this lead to the same signal/noise ratio as the localized stochastic sound source.

B. Results

We analyzed and resynthesized the stimuli \(x^{(L)}(n)\) and \(x^{(D)}(n)\) with the multichannel analysis/synthesis scheme described in Section III. We used 16 seconds of signal to compute the different estimators.

1) Sinusoidal analysis: To detect the sinusoids, we used the method described in Section III-B. We computed the STFT of the localized and diffuse signals with a 2-second long analysis window \(w_a\) (\(N_a = 96000\) samples at 48 kHz). We used the DPSW so that each sinusoidal component was represented by only \(K = 7\) bins, i.e., 3.5 Hz in the 96000-bin spectra. The frequency resolution was further increased by zero-padding \(w_a\) up to 8 seconds (\(N_a = 384000\) samples at 48 kHz) providing a spacing of 0.125 Hz between each frequency bin.

The average spectral estimator \(\hat{X}\) and the detected peaks are depicted on Figure 6. We set the peak detection threshold \(\delta\) to 11 dB. In both cases the two sinusoids were detected, with a frequency error of 0.023 and -0.012 Hz respectively. We can observe that, as predicted in Section III-B1, the variance of \(\hat{X}\) is reduced in the case of the diffuse sound field. At high frequencies (>1000 Hz) the \(Q = 80\) channels of \(x^{(D)}(n)\) are decorrelated. Consequently the variance of \(\hat{X}^{(D)}\) is reduced compared to \(\hat{X}^{(L)}\). The peak detection algorithm could benefit from this variance reduction to lower the threshold and detect sinusoidal components with lower signal/noise ratio. However, when analyzing a real aircraft sound, no assumption can be made on the spatial distribution of the noisy sources (localized or diffuse) so that the threshold must be set for the worst case (i.e., the localized case).
Due to the frequency estimation error, constant phase and amplitude deviations are introduced by our phase/amplitude estimation methods. More sophisticated phase estimation techniques could reduce the bias (e.g., by autocorrelation). With the DPSW zero-padded to 8 seconds, the amplitude bias is inferior to 0.2 dB. It could be further reduced by interpolating the spectral lobe of the analysis window, as proposed in [24]. However since the phase and amplitude bias are the same in all channels for each sinusoidal component, the interchannel phase and amplitude differences are respected, which is our only concern here. The amplitude and phase estimated in each channel for the sinusoids of the simulation are depicted on Figure 7.

After sinusoidal extraction, the residual was analysed and resynthesized with the method described in Section III-C. Note that to estimate the interchannel coherence and phase difference with a good accuracy, the analysis window should be longer than the maximum time delay between two microphones. With the simulated array, the maximum time delay was approximately 212 samples. For that reason, we used a 4096-sample analysis/synthesis window. After performing the deterministic and stochastic synthesis, both components were summed to get the full multichannel synthesized sound.

2) Spectral and spatial reconstruction: To assess the ability of the method to reconstruct the original spectral envelopes, interchannel coherences and phase differences, these three descriptors were computed on the original and synthesized sounds with a 4096-sample DPSW analysis window. Figure 8 illustrate the reconstruction of the three descriptors for both localized and diffuse source simulations. For the sake of readability, the descriptors are illustrated only for two representative channels: channel 1 which corresponds to a microphone located in one of the corners of the array, and channel 49 which is close to the center of the array. In the localized case, spectral envelopes were reconstructed with a maximum error (across all channel and all frequency bins) of 0.85 dB, a maximum IC error of 0.12 and a maximum IPD error of 0.86 rad. In the diffuse case, spectral envelope were reconstructed with a maximum error of 1.12 dB, a maximum IC error of 0.13 and a maximum IPD error of π rad. Please note that the IPD is significant and needs to be reproduced accurately only in frequency regions with high coherence. In the diffuse case, the coherence is high only for frequencies below approximately 200 Hz, and the IPD is well reconstructed in that region. These results confirm that the multichannel analysis/synthesis algorithm performs well for both localized and diffuse sources.

3) Evaluation of the model using sound field characterization: To further validate the efficiency of the multichannel analysis/synthesis algorithm, supplementary comparisons are provided. For that purpose, sound field extrapolation (SFE) and sound field characterization (SFC) methods as presented in [36], [37] are used. This SFC method is based on the computation of two physical metrics from the application of inverse problem theory to microphone array processing and sound field extrapolation [40]. The two scalar metrics are the energy vector magnitude ($R_E$, with $0 < R_E < 1$) and the directional diffusion ($D$, with $0 < D < 100\%$). To obtain these metrics, the original $x$ and synthesized $\hat{x}$ Q-channel signals are first converted to incoming plane wave representations in the frequency domain. Using discrete-time Fourier transform, $y_q(k)$ denotes the Fourier transform of $x_q(n)$. Accordingly, vectors $y(k) \in \mathbb{C}^{Q \times 1}$ and $\hat{y}(k) \in \mathbb{C}^{Q \times 1}$ will represent the original and synthesized signal for the frequency bin $k$. The plane wave representations are denoted $\mathbf{p}(k) \in \mathbb{C}^{L \times 1}$ and $\hat{\mathbf{p}}(k) \in \mathbb{C}^{L \times 1}$, respectively. They are both computed using a least-square approach with a Beamforming regularization matrix for each frequency bin $k$

$$
\mathbf{p} = \arg \min \{ ||\mathbf{Gp} - y||^2 + \lambda^2 ||\mathbf{Lp}||^2 \}
$$

where $L$ is the number of candidate plane wave sources, $\mathbf{G} \in \mathbb{C}^{Q \times L}$ is the matrix of complex amplitude of each $l$ plane wave to each $q$ microphone ($[\mathbf{G}]_{ql} = e^{-jk_ql}e^{j\phi_q}$), wavenumber vector $\mathbf{k}_l$ [rad/m] for propagation direction $l$ and the diagonal matrix $\mathbf{L} \in \mathbb{R}^{L \times L}$ is the beamforming regularization matrix given by $\mathbf{L} = [\text{diag} (||\mathbf{Gy}||/||\mathbf{Gy}||_{\infty})]^{-1}$. In eq 23, $\lambda$ is the regularization parameter that controls the
regularization amount. The solution of eq 23 is given by

$$p = \frac{G\gamma}{GG + \lambda^2 TL}$$  \hspace{1cm} (24)$$

Equations 23 and 24 are also computed for $\hat{p}$ to obtain the plane wave representation of the synthesized signals $\hat{p}$. For the evaluation reported in this paper, a set of $L = 642$ incoming plane waves (covering $4\pi$ steradians) was used and the processing was achieved for $Q$-channel one-second samples. Further explanations and developments that go beyond the scope of this paper can be found in Refs. [36], [37], [40]. The energy vector $E \in \mathbb{R}^{3 \times 1}$ is the normalized vector average of the incoming plane wave representations $p$ and $\hat{p}$ given by

$$E = \frac{\sum_{l=1}^{L} n_l |p_l|^2}{\sum_{l=1}^{L} |p_l|^2}$$  \hspace{1cm} (25)$$

and $E$ is the value of $\epsilon$ for a single impinging plane wave. In eq 27, $\langle E \rangle$ represents the spatial average of $E$ (i.e., $1/L \sum_{l=1}^{L} |E_l|$). The definition of the directional energy density $\mathcal{E} \in \mathbb{R}^{L \times 1}$ is given by

$$\mathcal{E}_l = \frac{|p_l|^2}{2\rho c^2}$$  \hspace{1cm} (28)$$

with air density $\rho$ [kg/m$^3$] and sound speed $c$ [m/s]. The directional energy density $\mathcal{E}$ represents the acoustical kinetic and potential energy densities coming from propagating directions $l$. Exact details and explanation of this metric are given in [36]. According to these definitions, $\mathcal{D} = 100\%$ for a perfectly diffuse sound field and $\mathcal{D} = 0\%$ for a single source in anechoic conditions.

In Refs. [36], [37], it was shown theoretically and experimentally that the combination of these two well-known metrics, the energy vector magnitude and directional diffusion, is able to distinguish archetypical sound fields: diffuse sound field, standing waves and directional sound field (i.e., one or few sources in free field). In this paper, the evaluation of the analysis/synthesis model using SFC is straightforward: 1) the SFC metrics are computed for the reference and synthesized signals ($R_E$, $\mathcal{D}$ and $\hat{R}_E$, $\mathcal{D}$, respectively) and 2) they are compared for several frequencies and situations (localized tone, localized noise and diffuse sound field, see Fig. 5). In [36], it was shown that the energy vector magnitude quantifies the directive character of the sound field, i.e., a large $R_E$ corresponds to one or few sources in free-field-like conditions while a small $R_E$ corresponds to a
standing wave pattern, a diffuse sound field or facing sound sources or sound reflections. In the same reference, it was verified that the directional diffusion $D$ is able to measure the global diffuse character of the sound field. It was also shown that, practically, for a finite microphone array, the directional diffusion, although theoretically bounded to 100%, rarely goes higher than 60%. The comparison of the SFC metrics are reported in Tabs. I and II where the relative difference (in %) is given by the difference of the reference and synthesis metrics normalized by the reference metrics ($100\% \times |R_E - \hat{R_E}|/R_E$ and $100\% \times |D - \hat{D}|/D$).

According to this comparison, the synthesized multichannel sounds share similar spatial metrics with the original multichannel sounds. More specifically, the energy vector magnitude and the directional diffusion are nearly perfectly recreated for the localized tone and noise. For the diffuse noises at 400 and 800 Hz, the directional diffusion is also well reproduced. However, the energy vector magnitude for the diffuse noises of the synthesized sounds differs from the reference sounds. This is caused by the fact that for low energy vector magnitude, i.e., for a sound that comes from facing directions, and more specifically for a diffuse-like sound field, the energy vector is highly dependent on the spatial polarity of the incoming sound field. In other words, a low energy vector magnitude tends to be much less precise. In all cases, the energy vector magnitude of the synthesized and reference sounds are well below the directive sound field threshold defined by $R_E \geq 0.6$ in [36], [37]. Therefore, in light of these observations, it is expected that the analysis/synthesis model will correctly reconstruct spatial features. In Tab. II, one also notes that the directional diffusion for the diffuse field at 800 Hz is low. This is caused by the fact that the corresponding wave length ($0.4288 \text{ m}$ with a sound speed of $343 \text{ m/s}$) is small in comparison with the overall size of the array. Therefore, 30 plane waves (see Fig. 5) are not enough to create a true diffuse sound field, in its strict sense [41], over the entire array region.

At 400 Hz, the diffuse sound is actually providing a larger directional diffusion since the wavelength is much larger ($1.7150 \text{ m}$ with a sound speed of $343 \text{ m/s}$) in comparison with the array region and the 30 decorrelated plane wave might therefore create an actual diffuse sound field over the entire array. These results are summarized in a graphical representation shown in Fig. 9 where one clearly notes that the localized sounds are well reconstructed in terms of these two spatial metrics. Moreover, the directional diffusion is well reconstructed for the diffuse-like sound fields.

As presented in [36], [37], sound-field scores are derived from these two metrics. These scores express the extent to which, according to the two metrics $R_E$ and $D$, the measured sound field is close to an ideal localized source, standing wave or diffuse sound field. The free-field score is given by

$$S_{ff} = R_E$$

The modal or standing-wave score is given by

$$S_m = (1 - R_E)(1 - D/60)^2$$

and the diffuse-field score is given by

$$S_d = (1 - R_E)(D/60)^2$$

The sound-field scores were computed both for the original and synthesized sound at the multichannel array. These scores are reported in Tabs. III and IV. Clearly, the synthesized multichannel signals share the same sound-field scores as the original. Previous remarks about the diffuse noise at 800 Hz also apply here and explain why both the reference and synthesized signals at 800 Hz for the diffuse signals made of 30 incoming uncorrelated plane waves are detected as a standing wave, i.e., the pattern is too regular over the array. These sound field scores are a supplementary demonstration of the efficiency of the proposed multichannel synthesis method to reproduce the general spatial character of the measured sound field.

## V. Potential Applications

The proposed model allows parametric transformations that would be difficult (if not impossible) to obtain based on signal processing techniques applied to raw aircraft recordings. The amplitude and frequency of sinusoidal components can be modified independently with a great accuracy. The noise spectral envelope, interchannel coherence and phase difference can also be manipulated, to reproduce different spatial configurations and control precisely sound field parameters like directivity/diffusion. These possibilities are of great interest for comfort studies in aircraft and vehicles.

Beside aircraft sound reproduction, analysis/synthesis of multichannel recordings finds many spatial sound applications. The proposed multichannel sinusoidal modeling offers a complementary approach to existing methods for sound source directivity measurement and reproduction [42], [43].

The model is suitable for manipulating the time-frequency and spatial characteristics of the synthesized sound field. Extending our method with short-time estimation of the spectral envelope, interchannel coherence and phase difference

<table>
<thead>
<tr>
<th>Source type</th>
<th>Frequency (Hz)</th>
<th>$R_E$ [0.1]</th>
<th>$\hat{R_E}$ [0.1]</th>
<th>Relative diff. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localized tone</td>
<td>999</td>
<td>0.985</td>
<td>0.984</td>
<td>0.016 %</td>
</tr>
<tr>
<td>Localized noise</td>
<td>800</td>
<td>0.981</td>
<td>0.981</td>
<td>0.016 %</td>
</tr>
<tr>
<td>Diffuse noise</td>
<td>800</td>
<td>0.164</td>
<td>0.218</td>
<td>33.056 %</td>
</tr>
<tr>
<td>Diffuse noise</td>
<td>400</td>
<td>0.281</td>
<td>0.174</td>
<td>38.262 %</td>
</tr>
</tbody>
</table>

Table I

**Comparison of the Energy Vector Magnitude for the Original $R_E$ and Synthesized $\hat{R_E}$ Multichannel Recordings.**

<table>
<thead>
<tr>
<th>Source type</th>
<th>Frequency (Hz)</th>
<th>$D$ (%)</th>
<th>$\hat{D}$ (%)</th>
<th>Relative diff. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localized tone</td>
<td>999</td>
<td>0.535 %</td>
<td>0.760 %</td>
<td>42.064 %</td>
</tr>
<tr>
<td>Localized noise</td>
<td>800</td>
<td>0.810 %</td>
<td>0.421 %</td>
<td>48.006 %</td>
</tr>
<tr>
<td>Diffuse noise</td>
<td>800</td>
<td>23.566 %</td>
<td>33.015 %</td>
<td>1.641 %</td>
</tr>
<tr>
<td>Diffuse noise</td>
<td>400</td>
<td>48.709 %</td>
<td>45.565 %</td>
<td>6.455 %</td>
</tr>
</tbody>
</table>

Table II

**Comparison of the Directional Diffusion for the Original $D$ and Synthesized $\hat{D}$ Multichannel Recordings.**


would potentially open these parametric transformations to any non-stationary musical or environmental sound.

A straightforward application example is spatial sound field synthesis and transformation, based on Rayleigh II integral [44], [45] as encountered in Wave Field Synthesis [44]. The Rayleigh II integral shows that an original sound pressure field can be uniquely recreated in a given volume $V$ if the sound pressure on its surrounding surface $S$ feed conformal dipole source distribution on $S$ with their main axis perpendicular to $S$. In this case one would make the analysis/synthesis of microphone array data over $S$ and play it back over a conformal dipole source array on $S$ to synthesize the sound field on one side of the reproduction source array. Since phase relationships between channels are correctly synthesized for tonal and correlated noise (see eqs. 11, 17 and 16), the sound field synthesis for these components should correspond to the original sound field in $V$. However, the synthesis interchannel uncorrelated noises are different random processes than the originals. Therefore, the array should reproduce a diffuse-like sound field with similar timbral characteristics as the original, but as a different random process. Depending on the shape of $V$, it would be possible to reproduce incoming sound fields (interior problem: wave field synthesis in a listening area surrounded by a loudspeaker array) or outgoing sound fields (exterior problem: directivity synthesis). Parametric resynthesis of the atomistic field components opens many potential applications, resulting in high-quality parametric time-frequency transformations (such as pitch-shifting, time-stretching, morphing) but also in parametric spatial transformations by manipulating individual sound field elements. Typical fictitious yet illustrative transformation examples would be: intensification of the diffuse sound field only, control of the diffusion amount, spatial rotation or translation of a tonal and localized sound components, etc.

Moreover, one should keep in mind that the method proposed in this paper can be applied to any stationary, vehicle-like, signal: microphone array data (such as for near-field acoustical holography [45]), binaural recordings, beamformer outputs in the time domain or loudspeakers tracks obtain from an arbitrary mixing or recording technique. In that sense, the proposed approach opens many potential applications within the field of spatial audio and sound localization and even sound imaging. Indeed, the capacity of the analysis/synthesis approach to separate the sinusoids from noise-like signals and to distinguish between correlated and uncorrelated noise is not yet applied, to our knowledge, to microphone array processing for sound source detection or sound field extrapolation.

### VI. Conclusion

A model for synthesizing multichannel aircraft sounds has been proposed. We presented a method for spectral and spatial analysis/synthesis of multichannel stationary sounds, collected with a microphone array or a binaural recording system. Sounds are represented as a collection of sinusoidal and noisy components, with controlled interchannel coherence and phase difference. The method was applied to a database of binaural sounds recorded in an aircraft cabin. Sound examples are available online [46]. A numerical simulation confirmed that the multichannel analysis/synthesis algorithm performs well for reconstructing stationary sinusoids with close frequencies, in presence of directive and diffuse noisy sources. Moreover, it was shown that the general spatial characteristics of the signals are well reconstructed according to two general sound field metrics. Our model allows real-time spectral and spatial

### Table III

<table>
<thead>
<tr>
<th>Source type</th>
<th>Frequency</th>
<th>$S_{Rf}$</th>
<th>$S_{m}$</th>
<th>$S_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localized tone</td>
<td>999</td>
<td>0.985</td>
<td>0.015</td>
<td>1.2316e-006</td>
</tr>
<tr>
<td>Localized noise</td>
<td>800</td>
<td>0.981</td>
<td>0.0189</td>
<td>3.4371e-006</td>
</tr>
<tr>
<td>Diffuse noise</td>
<td>800</td>
<td>0.164</td>
<td>0.574</td>
<td>0.262</td>
</tr>
<tr>
<td>Diffuse noise</td>
<td>400</td>
<td>0.282</td>
<td>0.245</td>
<td>0.473</td>
</tr>
</tbody>
</table>

Table III

**Sound field classification scores for the original sound field: free-field score $0 \leq S_{Rf} \leq 1$, standing-wave score $0 \leq S_{m} \leq 1$ and diffuse-field score $0 \leq S_d \leq 1$. Highest scores are shown in bold characters.**

### Table IV

<table>
<thead>
<tr>
<th>Source type</th>
<th>Frequency</th>
<th>$S_{Rf}$</th>
<th>$S_{m}$</th>
<th>$S_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Localized tone</td>
<td>999</td>
<td>0.984</td>
<td>0.016</td>
<td>2.5174e-006</td>
</tr>
<tr>
<td>Localized noise</td>
<td>800</td>
<td>0.981</td>
<td>0.019</td>
<td>9.560e-007</td>
</tr>
<tr>
<td>Diffuse noise</td>
<td>800</td>
<td>0.218</td>
<td>0.545</td>
<td>0.237</td>
</tr>
<tr>
<td>Diffuse noise</td>
<td>400</td>
<td>0.174</td>
<td>0.350</td>
<td>0.476</td>
</tr>
</tbody>
</table>

Table IV

**Sound field classification scores for the synthesized sound field: free-field score $0 \leq S_{Rf} \leq 1$, standing-wave score $0 \leq S_{m} \leq 1$ and diffuse-field score $0 \leq S_d \leq 1$. Highest scores are shown in bold characters.**

![Figure 9](image-url)

**Figure 9.** Sound field characterization results in the $R_{W}-D$ plane for the reference (○) and synthesized (+) signals. Black and red markers correspond to the localized tone at 999 Hz and the localized noise at 800 Hz. Blue and orange markers correspond to the diffuse field at 800 Hz (blue) and 400 Hz (orange).
modifications of the synthesis parameters for high-quality sound transformations. This work finds applications for aircraft simulators and comfort evaluation, and more generally for musical and environmental sound reproduction with multichannel systems.

VII. ACKNOWLEDGEMENTS

The authors would like to thank David Romblom and the anonymous reviewers for helpful comments on the manuscript.

REFERENCES


Charles Verron received a master’s degree in acoustics, signal processing and computer music from Université Paris 6 in 2004, and a Ph.D. degree in audio signal processing from Université de Provence in 2010 (Ph.D. realized at Orange Labs and CNRS-LMA, France). Between 2006 and 2012 he contributed to several audio projects at Computer and Audio Research Laboratory (Sydney University, Australia), Multimodal Interactions Laboratory (McGill University, Canada), REVES-INRIA (Sophia-Antipolis, France) and CNRS-LMA (Marseille, France). His main research interests include 3D audio and sound analysis/synthesis for games, music and virtual reality.

Philippe-Aubert Gauthier received a master’s degree in mechanical engineering from Université Laval in 2003 and a Ph.D. degree in mechanical engineering (sound field reproduction, best Ph.D. thesis award in sciences and engineering) from Université de Sherbrooke in 2007. He was a post-doctoral fellow at the Groupe d’Acoustique de l’Université de Sherbrooke where he worked on sound environment reproduction from 2007 to 2010. Since then, he is a professional researcher at Sherbrooke University. His main research interests include spatial sound, wave field synthesis, active sound control, sound localization. P.-A. Gauthier is also a musician and a professional sound artist.

Jennifer Langlois received a Ph.D. (2010) and a M.Sc. (2006) in Sensory Science from the Université de Bourgogne (France) and was initially trained in Biological Engineering in the Université de Lille (France). In her doctoral research, she used mixed-approaches to investigate various aspects of expertise development and tacit knowledge in the context of wine tasting. She is currently a post-doctoral research fellow at the Multimodal Interaction Laboratory and the Centre for Interdisciplinary Research in Music Media and Technology (Mcgill University, Canada), focusing on perceptual evaluation of audio synthesis and auditory comfort evaluation. Her research interests include sensory evaluation, verbal descriptors and cognitive aspects related to perception.

Catherine Guastavino is an Associate Professor at McGill University where she holds a William Dawson Research Chair. She received a Ph.D. in Psychoacoustics from the University of Paris (Pierre et Marie Curie) and post-doctoral training in cognitive psychology at McGill before joining the McGill School of Information Studies in 2005. She is a member of the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) where she served as Associate Director for Scientific and Technological Research from 2007 to 2009, and an Associate Member of the McGill Schulich School of Music. Her research interests include spatial audio, auditory localization, sound quality evaluations, soundscapes, and music perception, cognition and preservation.