ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 103 (2017) 1 – 1

Generating Sound Stimuli with Given Emergence Level and Low Frequency Content by Mixing Recordings

Ernesto Accolti[1], Federico Miyara[2], Fernando di Sciascio[1]

[1] Instituto de Automática, Universidad Nacional de San Juan - CONICET, Avda. San Martín 1109 oeste, San Juan, Argentina. eaccolti@inaut.unsj.edu.ar

[2] Laboratorio de Acústica y Electroacústica, Universidad Nacional de Rosario, Riobamba 245 bis, Rosario, Santa Fe, Argentina. fmiyara@fceia.unr.edu.ar

Summary

Noise annoyance and other effects of noise are reasonably correlated with A-weighted sound levels. Currently, the influence of low-frequency content and of sound emergence level (i.e. the sound level difference between total noise and residual noise) on effects of noise are being assessed in laboratories using stimuli based on recordings. Standards intended for regulations often include penalties depending on these factors. The difference between C- and A-weighted sound levels is frequently used as a descriptor of low frequency content. This work proposes a method to optimize the search and combination of a subset of audio files from a large set of noise sources recordings in order to achieve an environmental noise stimulus with previously specified values of factors (i) sound emergence level, (ii) C- A-weighted levels difference, (iii) A-weighted equivalent continuous sound level and (iv) duration of stimulus. The method is implemented, and one hundred stimuli are generated following a full factorial experimental design varying these four factors. The resulting stimuli show small differences between specified and measured values. The mean difference for factors i-iii is 0.46 dB and the maximum is 2.3 dB.

PACS no. 43.50.-x, 43.59.-e, 43.60.-c, 43.75.Tv, 43.75.Wx

1. Introduction

1.1. Soundscape generators

The present work proposes a realistic soundscape generation tool for the assessment of the relation of current standardized descriptors with effects of environmental noise on human beings. Similar works in the literature also propose tools for realistic soundscape generation [1, 2, 3, 4] or sounscape composition [5, 6].

One of these generators is intended for characterizing the mental representation of sound environments [1]. Others are intended for creating modern day *musique concrète* or acousmatic composition, sound design, soundscape composition, and other sonic sculpting tasks [5, 6]. A previous work [7] is also intended for the assessment of effects of environmental noise but related to parameters that are not standardized.

Possible applications of these generators are crowdsourcing [1] and immersive online media including virtual reality [2, 3] and augmented reality [2]. Applications on geographic issues are location-based sound search systems [4], sonic exploration [3] and sonification [2].

The outputs are generated by mixing recordings, modified recordings or resynthesis of recorded material from databases. The input data for these generators are of diverse nature. The inputs for the generator of Rossignol *et al.* [1] are high-level parameters for whole classes of sounds that are organized into a hierarchical semantically structured dataset. These high-level parameters are those of the statistical distributions of time intervals and gains applied to the mixed audio files. Several of these generators manually mix recordings based on analysis , transformation, and resynthesis of natural sounds [2, 5].

The inputs for generator of Misra *et al.* [5] are the controls of tools for time-stretch, shrunk or pitch-shift, periodicity, density and randomness. Other input methods are specifying a travel or a geography [2, 3] and natural language queries that can be sent by several users [6]. A recent generator use acoustic parameters such as global spectra and the density of sound events depending on their duration or semantic category [7].

Current laboratory research on effects of noise such as annoyance potential are conducted with auralization techniques in simulated environments. In order to achieve realism on stimuli, audio files previously recorded from real sound sources are properly combined [7, 8, 9]. The realism is a property quite related to the sense of presence. Finney and Janer [2] adopted questions relevant for soundscape

Received 29 December 2016,
accepted 4 July 2017.

generators from a presence questionnaire [10] developed for virtual environments.

1.2. Environmental noise parameters

Environmental noise caused by several sources is defined [11] in terms of *total noise*, *residual noise* and *specific noise* for measurement and assessment purposes. Total noise is composed by the specific noises mixed together with the residual noise.

Noise regulations and standards on environmental noise often specify limits in terms of A-weighted equivalent continuous sound level $L_{\mathrm{A,eq}}$. Sound emergence level $L_{\mathrm{A\text{-}R}}$ is defined as the difference between total noise and residual noise levels [11, 12]. In several cases, as when existing noise levels are low, the limits also include sound emergence level. Standards [13, 14] use a procedure that is equivalent to a limit on sound emergence level.

Alayrac *et al.* [8] enumerate some studies that found significant effects of residual noise on annoyance caused by aircraft noise and also enumerate some other studies that do not support this significance. Finally they report experimental evidence of the effects of the type of residual noise, the type of industrial noise source and its sound emergence level on total annoyance.

The difference between the C- and A-weighted equivalent continuous levels ($L_{\mathrm{C\text{-}A}}$) is used as a measure of low-frequency content [9, 14, 15, 16, 17, 18]. If low-frequency content exceeds certain limit, a penalization is applied to $L_{\mathrm{A,eq}}$ [14, 16]. Kjellberg *et al.* [17] found a small but significant effect of $L_{\mathrm{C\text{-}A}}$.

1.3. Purpose and objectives

The purpose of this article is to introduce a method to combine signals from an audio files database in such a way that a simulated environmental noise with previously specified values for $L_{\mathrm{A,eq}}$, $L_{\mathrm{C\text{-}A}}$ and $L_{\mathrm{A\text{-}R}}$ is achieved. The descriptors correspond to the listening position of a virtual scenario simulated with auralization techniques. The intended use of the so-obtained signal is to investigate the effect of $L_{\mathrm{C\text{-}A}}$ and $L_{\mathrm{A\text{-}R}}$ on noise annoyance potential. The method proposes a complement for on-going research about effects of noise.

In Section 2 the problem is formulated as a mathematical optimization problem complemented with strategies for the temporal placement of recordings. Then in Section 3 an hypothetical scenario is configured and the implementation details for the optimization formulation and for the temporal placement of recordings is shown. In Section 4 the method is evaluated composing a set of 100 stimuli. The errors of this set and a subjective evaluation of a subset of these stimuli are reported and analyzed. Finally, Sec. 5 contains the conclusions and possible future work.

2. Stimuli composition method

The mathematical formulation of the proposed method is shown in this section. The main task is to compose an audio file, containing an aural stimulus with previously specified values of $L_{\mathrm{A,eq}}$, $L_{\mathrm{C\text{-}A}}$ and $L_{\mathrm{A\text{-}R}}$ descriptors. The output file is composed combining a subset of audio files from a database.

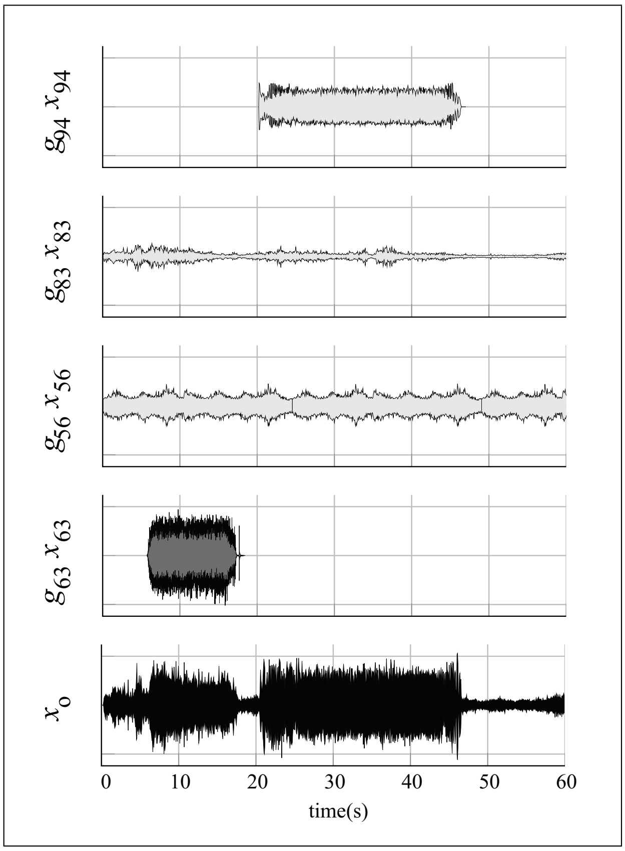


Figure 1. Mixing process example. Each row contains a time series. The first three rows are (filled light gray) the files that compose the residual noise affected by their corresponding gains. First row: x_{94} is an engine sound. Second row: x_{83} is a wind sound. Third row: x_{56} is a sea sound. Forth row (filled dark gray): specific noise under investigation, x_{63}, a sewing machine. Each file is shown affected by their corresponding gains g_i. The fifth row (filled black) is the signal of the output file x_{o}.

We formulate the audio combination as a linear combination of signals corresponding to a selected subset of audio files from a database. The coefficients of the audio combination are gains applied to each audio file from the selected subset. Figure 1 shows an example of the combination of the 4 signals of the first rows into the signal of the last row.

We formulate an optimization problem to find the adequate subset of files and the corresponding gains in order to reach the previously specified values of $L_{\mathrm{A,eq}}$, $L_{\mathrm{C\text{-}A}}$ and $L_{\mathrm{A\text{-}R}}$ when auralizing the mixed signal. Before the audio combination, we automated a time shift for each audio file from the selected subset. These time shifts follow a strategy to obtain a realistic distribution of the sound events that the output file will contain.

2.1. Audio Combination

Each stimulus is achieved by mixing a subset of audio files from a database of N audio files containing the signals $x_1, x_2, \ldots, x_N$. The mixing process involves the linear combination of the temporally located signals corresponding to each audio file from the selected index subset $I \subseteq \{1, \ldots, N\}$. Each file x_i may be repeated Q_i times. The output file $x_\mathrm{o}(t)$, of duration T_o, is, for $0 \le t \le T_\mathrm{o}$,

$$x_\mathrm{o}(t) = \sum_{i \in I} g_i \sum_{q=0}^{Q_i} x_i(t - t_{i,q}), \tag{1}$$

where i are the index from the selected subset I, g_i are corresponding gains, $t_{i,q}$ are the corresponding insertion instants and q the index of repetition.

In the example of Figure 1, the selected subset is $I = \{56, 63, 83, 94\}$. The specific noise, x_{63}, is a noise from a sewing machine. The residual noise is composed by a sea sound x_{56}, a wind sound x_{83} and an engine sound x_{94}. Total noise is estimated by (1) as

$$\begin{aligned} x_\mathrm{o}(t) \;=\; & g_{56}x_{56}(t - t_{56,0}) + g_{56}x_{56}(t - t_{56,1}) \\ & + g_{56}x_{56}(t - t_{56,2}) + g_{63}x_{63}(t - t_{63,0}) \\ & + g_{83}x_{83}(t - t_{83,0}) + g_{94}x_{94}(t - t_{94,0}), \end{aligned} \tag{2}$$

for $0\,\mathrm{s} \le t \le 60\,\mathrm{s}$.

File x_{56} is inserted 3 times. The first time x_{56} is placed at $t_{56,0}$, the first repetition is placed at $t_{56,1}$ and the second repetition at $t_{56,2}$.

The file used as the specific noise [19] and the ones used to compose the residual noise [20, 21, 22] are available in the web.

Section 2.2 formulates a linear programming problem (MILP formulation as will be addressed in Section 2.2.4) in order to find both the selected subset I and the gains g_i which are required for the mixing process. The expected sound exposures are defined from the values of $L_\mathrm{A,eq}$, $L_\mathrm{C\text{-}A}$, $L_\mathrm{A\text{-}R}$ and T_o that should all be specified in order to request the generation of a stimulus (See Section 2.2.1). The problem is formulated as the combination of the sound exposures of the individual audio files from the database (See Section 2.2.2) subject to constraints related to realism and the definition of the problem (See Section 2.2.3).

In Section 2.3 the temporal composition is briefly described using a state-of-the-art method. The output of this composition are the insertion instants $t_{i,q}$ corresponding to each audio file x_i and to each repetition index $q \in \{0, 1, \ldots, Q_i\}$. Each audio file x_i is repeated Q_i times. When $Q = 0$ audio file is inserted only one time (i.e. it is not repeated any time) and $q \in \{0\}$.

2.2. Exposure combination problem

Let the sound database be defined as $X = \{x_1, x_2, \ldots, x_N\}$, a set of N audio files, each of them of duration T_n. The problem is to find an index subset $I \subseteq \{1, \ldots, N\}$, an index subset $J \subset I$ and, for each $i \in I$, a gain $g_i \in \mathbb{R}^+$ such that, when the files x_i are mixed together into an output file x_o and reproduced into a virtual scenario we get an environmental noise whose values for $L_\mathrm{A,eq}$, $L_\mathrm{C\text{-}A}$ and $L_\mathrm{A\text{-}R}$ descriptos were previously specified.

The residual noise x_r could be generated mixing only files x_j with $j \in J$. The audio files that represent the specific noise under investigation are identified with the index subset $K = I - J$ (i.e. the subset of files present in the total noise and not in the residual noise). In the example of Figure 1 the selected subsets are $I = \{56, 63, 83, 94\}$, $J = \{56, 83, 94\}$ and $K = \{63\}$.

The mixing process consists in summing the temporally placed signals x_i affected by gains g_i (or x_j affected by gains g_j in the case of residual noise). In the example of Figure 1 the residual noise could be generated mixing only the signals of the three first rows (the signals filled with light gray).

2.2.1. Specified instance

Let the A-weighted sound exposure be

$$\begin{aligned} e_\mathrm{A} &= \int_0^{T_\mathrm{o}} p_\mathrm{A}^2(t)dt && (a) \\ &= p_{\mathrm{A},rms}^2 \times T_\mathrm{o} && (b) \\ &= p_0^2 \times 10^{L_\mathrm{A,eq}/10} \times T_\mathrm{o}, && (c) \end{aligned} \tag{3}$$

where the integration is computed over the duration T_o of audio file x_o, $p_0 = 20\ \mu\mathrm{Pa}$ is the reference sound pressure, $p_\mathrm{A}(t)$ is the A-weighted sound pressure due to the output file x_o measured at the position of the subject and $p_{\mathrm{A},rms}$ its root-mean-square value. The sound pressure due to x_o, as measured at the position of the subject, includes effects of D/A conversion, amplification, transduction and propagation effects (Details will be addressed in Section 2.4).

Similarly, e_R is the A-weighted exposure to residual noise (i.e. due to x_r). And e_C is the expected exposure to total noise (i.e. due to x_o) but using C-weighting instead of A-weighting network.

The problem can be defined for several instances. Each instance should specify the exposure values for e_A, e_C and e_R. Notice that

$$e_\mathrm{A} = p_0^2 \times 10^{\frac{L_\mathrm{A,eq}}{10}} \times T_\mathrm{o} \tag{4}$$

$$e_\mathrm{C} = p_0^2 \times 10^{\frac{L_\mathrm{A,eq} + L_\mathrm{C\text{-}A}}{10}} \times T_\mathrm{o} \tag{5}$$

$$e_\mathrm{R} = p_0^2 \times 10^{\frac{L_\mathrm{A,eq} - L_\mathrm{A\text{-}R}}{10}} \times T_\mathrm{o} \tag{6}$$

Thus, from (4-6), the values for $L_\mathrm{C\text{-}A}$, $L_\mathrm{A\text{-}R}$, $L_\mathrm{A,eq}$ and T_o are used to define an instance.

2.2.2. Exposure combination

The proposed audio combination involves the sequential addition, the simultaneous addition and the partially overlapping addition of individual signals. Sound exposure is proportional to the energy of the signal in an audio file. The energy of a given signal composed as the sequential addition of individual signals can be computed by the sum of the energies of these individual signal. This computation scheme also applies when individual signals are added simultaneously if they are mutually incoherent. Furthermore, the scheme also applies when individual signals partially overlap (overlapping parts should satisfy the mutual incoherence condition). This calculation scheme is extended below for the linear combination of exposures.

Let $e_{\mathrm{A},n}$ and $e_{\mathrm{C},n}$ be, respectively, the A-weighted and the C-weighted exposure due to audio file x_n. The integration period matches its audio duration T_n. The squared gain coefficients w_n for the total noise are defined as

$$w_n = \begin{cases} g_n^2 & \text{if } n \in I \\ 0 & \text{otherwise}, \end{cases} \tag{7}$$

while the squared gain coefficients v_n for the residual noise are defined as

$$v_n = \begin{cases} g_n^2 & \text{if } n \in J \\ 0 & \text{otherwise}, \end{cases} \tag{8}$$

where g_n is the gain that should be applied to audio file x_n in the mixing process to generate both output files x_o and x_r. These definitions consider squared gains for simplicity because sound exposure is proportional to the squared sound pressure as shown in (3).

Assuming incoherence between audio files x_i with $i \in I$, the A-weighted exposure to the total noise is

$$e_\mathrm{A} = e_{\mathrm{A},1} w_1 + e_{\mathrm{A},2} w_2 + \cdots + e_{\mathrm{A},N} w_N, \tag{9}$$

and the C-weighted exposure to the total noise is

$$e_\mathrm{C} = e_{\mathrm{C},1} w_1 + e_{\mathrm{C},2} w_2 + \cdots + e_{\mathrm{C},N} w_N. \tag{10}$$

Assuming incoherence between audio files x_j with $j \in J$, the A-weighted exposure to the residual noise is

$$e_\mathrm{R} = e_{\mathrm{A},1} v_1 + e_{\mathrm{A},2} v_2 + \cdots + e_{\mathrm{A},N} v_N. \tag{11}$$

Notice that assuming incoherence, the summation of exposures of (9-11) does not depend on whether the audio signals x_n are simultaneous or not.

2.2.3. Constraints

Let $\mathbf{1}_I : \{1, 2, \ldots, N\} \to \{0, 1\}$ be the indicator function of subset I that indicates if audio file x_n is included in the total noise. And let $\mathbf{1}_J : \{1, 2, \ldots, N\} \to \{0, 1\}$ be the indicator function of subset J that indicates if audio file x_n is included in the residual noise. The constraint that audio files included in residual noise should be included in the total noise (i.e. $J \subset I$) can be modeled in terms of the indicator functions as

$$\mathbf{1}_J(n) \leq \mathbf{1}_I(n). \tag{12}$$

Let $\mathbf{1}_C : \{1, 2, \ldots, N\} \to \{0, 1\}$ be the characteristic function of subset C containing the indexes n such that x_n may be a specific noise under investigation (e.g. in a context of the assessment of annoyance caused by industrial noise, the natural sounds are not considered as possible specific noises). The number of audio files that simulate the specific noises under investigation is the cardinality of subset K ($\#K$). Note that $\mathbf{1}_C(n) = 1$ does not imply that x_n is actually a specific noise (i.e. does not imply $n \in K$) because it can be part of the residual noise (i.e. $n \in J$) or even not in the total noise (i.e. $n \notin I$), depending on the solution of the whole problem. Let $\#I_\mathrm{C} = \sum \mathbf{1}_I(n) \times \mathbf{1}_C(n)$ be the number of noises with $n \in C$ that are present in the total noise, and let $\#J_\mathrm{C} = \sum \mathbf{1}_J(n) \times \mathbf{1}_C(n)$ be the number of potential specific noises that are present in the residual noise. The following constraint models the number of audio files that simulate the specific noises under investigation

$$\#I_\mathrm{C} - \#J_\mathrm{C} = \#K. \tag{13}$$

An upper bound w_n^{max} should be defined to avoid excessive amplification that could be perceived as a source placed closer than usual. This bound accounts that $w_n = 0$ for $n \notin I$ because of the definition of w_n given in (7)

$$w_n \leq \mathbf{1}_I(n) \times w_n^{max}. \tag{14}$$

A lower bound w_n^{min} should be used to prevent sound events from being masked by other sound events in the file x_o. This bound also accounts that $w_n = 0$ for $n \notin I$ because of (7),

$$w_n \geq \mathbf{1}_I(n) \times w_n^{min}. \tag{15}$$

Similar upper and lower bounds for v_n define the following constraints,

$$v_n \leq \mathbf{1}_J(n) \times v_n^{max}, \tag{16}$$

$$v_n \geq \mathbf{1}_J(n) \times v_n^{min}. \tag{17}$$

The definitions of w_n and v_n, from equations (7) and (8), imply that the value of v_n should equal w_n for $n \in J$ because $J \subset I$. The following condition,

$$v_n = \begin{cases} w_n, & \text{if } \mathbf{1}_J(n) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

could be modified into the following linear form,

$$\begin{aligned} v_n - w_n &\leq M(1 - \mathbf{1}_J(n)), \\ w_n - v_n &\leq M(1 - \mathbf{1}_J(n)), \end{aligned} \tag{19}$$

where M is a sufficiently large number in order to ensure that constraint of w_n is not tight (i.e. $M \geq w_n^{max}$). These constraints ensure that $w_n = v_n$ when $\mathbf{1}_J(n) = 1$.

2.2.4. Optimization formulation

The problem can be addressed using various optimization techniques such as least squares, linear programing, and subclasses of linear programming such as integer programing. Particularly, this work addresses the problem using a mixed integer linear programming (MILP) formulation [23].

Let $\mathbf{E}_\mathrm{A} \in \mathbb{R}^{1 \times N}$ be a row vector whose components are the A-weighted exposure $e_{\mathrm{A},n}$ corresponding to each audio file x_n from the database

$$\mathbf{E}_\mathrm{A} = \left[e_{\mathrm{A},1} \; e_{\mathrm{A},2} \; \cdots \; e_{\mathrm{A},N} \right], \tag{20}$$

and $\mathbf{E}_\mathrm{C} \in \mathbb{R}^{1\times N}$ be an other row vector whose components are the C-weighted exposure $e_{\mathrm{C},n}$ corresponding to each audio file x_n from the database

$$\mathbf{E}_\mathrm{C} = \left[e_{\mathrm{C},1}\ e_{\mathrm{C},2}\ \cdots\ e_{\mathrm{C},N}\right]. \tag{21}$$

Then we define the exposure matrix $\mathbf{A} \in \mathbb{R}^{3\times 2N}$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{E}_\mathrm{A} & \mathbf{0}_{(\mathbf{N})} \\ \mathbf{0}_{(\mathbf{N})} & \mathbf{E}_\mathrm{A} \\ \mathbf{0}_{(\mathbf{N})} & \mathbf{E}_\mathrm{C} \end{bmatrix}, \tag{22}$$

where $\mathbf{0}_{(\mathbf{N})}$ are N sized row vectors of null elements.

Let $\vec{c} = [v_1, \cdots, v_N, w_1, \cdots, w_N]^T$ be the squared gain coefficients vector and $\mathbf{u} = [e_\mathrm{R}, e_\mathrm{A}, e_\mathrm{C}]^T$ be an instance vector containing the values of exposure required to the output files.

The linear combinations of exposures from equations (9), (10) and (11) can be incorporated into the matrix form

$$\mathbf{A}\vec{c} = \mathbf{u}. \tag{23}$$

Denote the approximate solution as $\hat{\mathbf{u}} = [\hat{e}_\mathrm{R}, \hat{e}_\mathrm{A}, \hat{e}_\mathrm{C}]^T$. The cost function is defined in terms of the error $\hat{\mathbf{u}} - \mathbf{u}$. We minimize the elements $|\hat{e}_\mathrm{R} - e_\mathrm{R}|$, $|\hat{e}_\mathrm{A} - e_\mathrm{A}|$ and $|\hat{e}_\mathrm{C} - e_\mathrm{C}|$ by minimizing the cost function $f_s = \|A \times \vec{c} - \vec{u}\|_\infty$ subject to constraints from Section 2.2.3. The Chebyshev distance, or ℓ_∞ norm, is proposed as the cost function.

An ℓ_∞ linearization [23] is applied below in order to approximate the cost function f_s. Let $z \in \mathbb{R}_{\geq 0}$ be an auxiliary variable and let $f_l = z$ be the new linear cost function to be minimized. Then the ℓ_∞ linearization is formulated as the following problem.

$$\begin{array}{lll} \text{minimize} & z & \\ \text{subject to} & \mathbf{A}\vec{c} - \vec{u} \leq z\mathbf{1}, & \text{(a)} \\ & \mathbf{A}\vec{c} - \vec{u} \geq -z\mathbf{1}, & \text{(b)} \\ & \mathbf{1}_J(n) \leq \mathbf{1}_I(n), & \text{(c)} \\ & \#I_\mathrm{C} - \#J_\mathrm{C} = \#K, & \text{(d)} \\ & w_n \leq \mathbf{1}_I(n) \times w_n^{max}, & \text{(e)} \\ & w_n \geq \mathbf{1}_I(n) \times w_n^{min}, & \text{(f)} \\ & v_n \leq \mathbf{1}_J(n) \times v_n^{max}, & \text{(g)} \\ & v_n \geq \mathbf{1}_J(n) \times v_n^{min}, & \text{(h)} \\ & v_n - w_n \leq M\left[1 - \mathbf{1}_J(n)\right], & \text{(i)} \\ & w_n - v_n \leq M\left[1 - \mathbf{1}_J(n)\right], & \text{(j)} \end{array} \tag{24}$$

where $n = 1, 2, \ldots, N$ and $\mathbf{1}$ is a vector with 3 elements equal to the unity.

The constraints (24.a) and (24.b) are introduced to approximate the ℓ_∞ norm in a linear form. Notice that minimizing $f_l = z$ subject to (24.a) and (24.b) is equivalent to minimize the maximum of the elements of $\hat{\mathbf{u}} - \mathbf{u} = (\hat{e}_\mathrm{R} - e_\mathrm{R}, \hat{e}_\mathrm{A} - e_\mathrm{A}, \hat{e}_\mathrm{C} - e_\mathrm{C})$, in turn equivalent to the definition of an ℓ_∞ norm. Constraint (24.d) restricts the problem to the equation (13). Constraints (24.c) and (24.e-j) restrict the problem to the inequations (12), (14), (15), (17), (17) and (19), respectively.

For the formulation as a mixed integer linear programming problem the objective variables are: the elements of the squared gain coefficients vector $\vec{c} \in \mathbb{R}^{2N}_{\geq 0}$; the indicator functions ($\in \{0, 1\}^{2N}$ corresponding to $\mathbf{1}_I(n)$ and $\mathbf{1}_J(n)$ with $n \in \{1, 2, \ldots, N\}$); and the auxiliary variable $z \in \mathbb{R}_{\geq 0}$. In Section 3.3 the objective variables are concatenated in a vector $\vec{b}$ for implementation purposes.

2.3. Temporal distribution of sound events

Once the optimization problem has been already solved, the mixing process inputs are: the set of audio files I, the gains g_i for each $i \in I$, and the insertion instants for each audio file x_i with $i \in I$. The set I and gains g_i are found solving the problem stated in (24). The insertion instants $t_{i,q}$ are computed using a method similar to the one presented in a previous work [7]. The method consist in two calculation schemes based on two strategies called *loop strategy* and *Poisson process strategy*. Audio files containing long stationary noises are placed following the loop strategy and files containing single sound events or groups of sound events are placed following the Poisson process strategy.

Audio files are edited when including them into the sound database in order to contain (i) an individual sound event when it sounds realistic distributed as a Poisson process, (ii) a group of similar events when Poisson process is not realistic or the temporal distribution is unknown or (iii) a long stationary noise when a continuous noise is the most realistic approach.

The Poisson process strategy consists in finding insertion instants $t_{i,q}$ following a Poisson distribution. Doing so for audio files that contain group of events, the time intervals between events will internally follow the recorded distribution and not necessarily a Poisson distribution. Whereas for audio files containing only one event, they will follow a Poisson distribution with respect to events of other audio files of the same group.

The loop strategy, used for audio files containing long stationary noises that are seamlessly loopable, consists in placing the audio file x_i throughout the whole duration of output file x_o. The loops are automated during the mixing process. The audio file x_i is looped, trimmed, or both, depending on the duration T_i and the specified duration T_o. In case $T_i \geq T_\mathrm{o}$ the index $q \in \{0\}$ indicating that file x_i is placed only one time. Otherwise, q indicate the repetition number $q \in \{0, 1, \ldots, Q\}$, and the number of repetitions is

$$Q = T_\mathrm{o} \setminus T_i, \tag{25}$$

where $\setminus$ is an integer division (i.e. a division where the remainder is discarded).

A simple modification is applied to the exposure of files that could be looped during the mixing process. It consists in scaling the exposure $e_{\mathrm{A},n}$ and $e_{\mathrm{C},n}$ before the definition of $\mathbf{E}_\mathrm{A}$ in (20) and $\mathbf{E}_\mathrm{C}$ in (21) only for those flies with long stationary noises. The scale factor is T_o/T_n in order to account for the fact that the exposure duration of a file x_n looped over the duration of audio file x_o is T_o.

In this work we allow repetitions only for the loop strategy. Thus, for files containing a single sound event or groups of sound events, repetitions are not allowed (i.e. $q \in \{0\}$).

2.4. Audio and acoustic signals

Effects of D/A conversion, amplification, transduction and propagation should be considered in order to model the relationship of digital audio signals with sound pressure, for a given auralization configuration.

These effects can be considered in two parts. The first part includes the real effects from the digital signal to the subject's ear (e.g. D/A conversion, the real room response, the response of the headphone housing or the sound isolation of a partition in case the loudspeaker is placed behind a real partition). The second part are simulated effects. Simulated effects can be addressed before D/A conversion to simulate particular environments (e.g. simulated propagation effects, isolation of a virtual façade or virtual room response) and to compensate some effects of the first part (e.g. the free- or diffuse-field response compensation for audiometric headphones or inverse transfer function of loudspeakers).

Assuming D/A conversion, amplification, transduction and propagation effects can be modeled as linear time-invariant (LTI) systems, the A-weighted sound pressure used in (3) to compute the A-weighted exposure can be computed as

$$p_{\mathrm{A}}(t) = \big(x_{\mathrm{o}} * h_{\mathrm{A}} * h_{\mathrm{s}} * h_{\mathrm{o}}\big)(t), \tag{26}$$

where $h_{\mathrm{o}}(t)$ is the impulse response that integrates the effects of D/A conversion, amplification, transduction and real propagation effects, h_{s} is the impulse response of simulated effects, $h_{\mathrm{A}}(t)$ is the impulse response of the A-weighting network and $x * h$ denotes the convolution of x with h. Similarly the C-weighted sound pressure can be computed with (26) using the impulse response of the C-weighting network $h_{\mathrm{C}}(t)$, instead of $h_{\mathrm{A}}(t)$.

Sound exposure $e_{\mathrm{A},n}$ (and $e_{\mathrm{C},n}$) due to x_n should be estimated at the position of the subject and after simulated effects, effects of D/A conversion, amplification, transduction and propagation. The sound exposure for each audio file x_n can be estimated directly by measurement or by modeling the effects of the given auralization configuration. Mathematically, assuming LTI systems, both estimations can be expressed as

$$e_{\mathrm{A},n} = \int_0^{T_n} \big[\big(x_n * h_{\mathrm{A}} * h_{\mathrm{s}} * h_{\mathrm{o}}\big)(t)\big]^2 \mathrm{d}t \tag{27}$$

and

$$e_{\mathrm{C},n} = \int_0^{T_n} \big[\big(x_n * h_{\mathrm{C}} * h_{\mathrm{s}} * h_{\mathrm{o}}\big)(t)\big]^2 \mathrm{d}t. \tag{28}$$

3. Implementation

In this section the implementation details of the linear programming problem, the temporal placement strategies and the mixing process formulated in Section 2 are shown. The implementation includes the definition of the audio database and its associated meta-data (Sec. 3.1), the configuration of the scenario (Section 3.2), the preparation of the input data for the solver and the determination of constants that complete the optimization problem (Section 3.3). Finally, the automated audio mixing process is shown following an example used to compose a stimulus (Section 3.4).

3.1. Sound database

A sound database of $N = 97$ audio files is used for this implementation. In order to include an audio file to the database a set of interactive operations are carried out. A second set of automated tasks complete the inclusion of a file or a set of files in the database.

The interactive operations include (i) recording or finding sounds from other databases, (ii) defining a calibration constant (i.e. the sound pressure corresponding to the full scale of the audio file), (iii) defining the temporal category for the audio file (i.e. single events, groups of events or long stationary noises), (iv) audio editing to isolate the desired events and avoid clicks (Section 2.3) and (v) indicate if file x_n may potentially be a specific noise $\mathbf{1}_C(n) = 1$ or not $\mathbf{1}_C(n) = 0$ (Sec. 2.2.3). Sounds from free databases [24, 25] and self made recordings were used in this implementation.

The automated operations include computing for each file its duration T_n and its $1/3$ octave bands spectrum $\{L_{e,1,n}, \dots L_{\mathrm{e},m,n}, \dots L_{e,24,n}\}$ from the band $m = 1$ centered at 50 Hz to the band $m = 24$ centered at 10 kHz. The estimated values are entered in a meta-data file. Each entry in the meta-data file, for each audio file x_n, includes its calibration constant, its temporal category, its duration T_n, its indicator as potentially specific noise $\mathbf{1}_C(n)$, and its spectrum $\{L_{e,1,n}, \dots L_{\mathrm{e},m,n}, \dots L_{e,24,n}\}$.

3.2. Scenario configuration

In order to test the methodology generating a set of stimuli we configured a hybrid scenario with real and simulated components. We determined the impulse response of the real part $h_{\mathrm{o}}(t)$ and simulated the impulse response of a partition as $h_{\mathrm{s}}(t)$. Thus, the exposures at the position of subject in this scenario are computed using (26) and (3), in order to analyze how close the specified values of exposure levels $L_{\mathrm{A,eq}}$ and $L_{\mathrm{C,eq}}$ are reached due to audio file x_{o} at the position of the subject for a given instance.

A loudspeaker was placed in a room of 7 m long, 4 m wide and 3 m high with sound absorption in the walls. The loudspeaker is an Alessis M1 Active MKII, hidden by a thin and opaque curtain. Figure 2 shows a diagram of the room including the source position (top right corner). The room floor is made of ceramic tiles and the ceiling, of plaster. The 70 % of wall surfaces are cladded with a product of

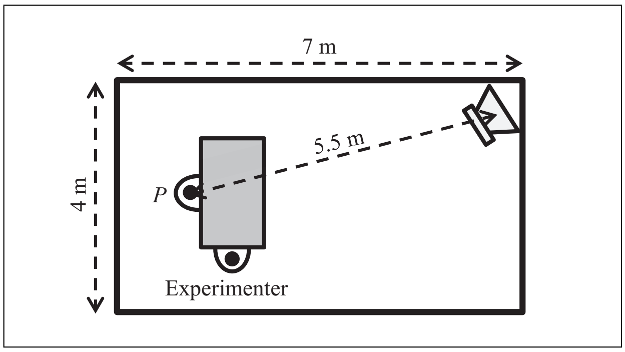


Figure 2. Listening room and position of the source and the microphone.

dense mineral fiber and the remaining 30 % are wood panels. The room has a table and five chairs. The impulse response $h_o(t)$ was identified using ISO 18233 method [26] at a microphone placed in position P at 1.2 m high and more than 1 m from any reflecting surface. The reverberation time of the room is 0.75 s and the background noise is below a Noise Criteria NC 20.

In order to simplify (27) and (28) we implemented the time convolution as the summation of exposure levels and attenuation for real and simulated effects in 1/3 octave bands. Third octave band attenuation levels from $h_o(t)$ and $h_s(t)$ were estimated as $A_{o,m}$ and $A_{s,m}$, respectively. Then (27) is estimated as

$$e_{A,n} = \sum_{m=1}^{24} 10^{\frac{L_{e,m,n}-A_{o,m}-A_{s,m}+C_A}{10}}, \tag{29}$$

where $C_{A,m}$ is the 1/3-octave band A-weighting correction for band m.

And (28) is estimated as

$$e_{C,n} = \sum_{m=1}^{24} 10^{\frac{L_{e,m,n}-A_{o,m}-A_{s,m}+C_C}{10}}, \tag{30}$$

where $C_{C,m}$ is the 1/3-octave band C-weighting correction for band m.

The calculation of the convolution in (27) and (28) can be implemented by the Fast Fourier Transform technique. This technique takes approximately 6×10^5 operations for each frame of 40 000 samples of each x_n and h_o. The proposed simplification significantly reduces the computation load each time a new scenario is proposed. Only 72 operations are used to calculate (29) or (30) for the total length of each file. The simplified technique does not require access to the audio signal itself but only the spectrum from the corresponding file entry in the meta-data file. Thus, a great reduction on the computation load is achieved by implementing this simplification.

3.3. Optimization

We implemented the problem in (24) using the MILP package of CPLEX Solver [27]. CPLEX is a state-of-the-art mathematical programming tool designed for the resolution of Mixed Integer Linear Programming problems (among other problems). This solver, and similar solvers of the same purpose, require as inputs:

- a vector containing the coefficients of a linear cost function
- a matrix and a vector of inequalities
- a vector indicating if each objective variable is continuous, integer or Boolean
- (optionally) a matrix and a vector of equalities
- (optionally) a lower bounds vector
- (optionally) an upper bounds vector
- (optionally) a starting point

The main output is a vector containing the solution with the form of the objective variable. Other outputs include the evaluation of the cost function and details about the performance of the solver algorithm.

3.3.1. Objective variables and cost function

For the implementation purposes, we grouped the objective variable in a vector $\vec{b}_{(4N+1)}$

$$\begin{aligned}\vec{b} = \big[&v_1, \dots, v_N, w_1, \dots, w_N, \\ &\mathbf{1}_J(1), \dots, \mathbf{1}_J(N), \\ &\mathbf{1}_I(1), \dots, \mathbf{1}_I(N), z\big]^{\mathrm{T}}.\end{aligned} \tag{31}$$

The cost function is introduced to the solver as a vector $\mathbf{f}$ whose elements are the coefficients of the cost function $f_s = \mathbf{f} \cdot \vec{b}$. In our case, in order to obtain $f_s = z$ as in (24), we define $\mathbf{f} = [\mathbf{0}_{(\mathbf{4N})},\ 1]^T$, where $\mathbf{0}_{(\mathbf{4N})}$ is a $1 \times 4N$-sized row vector of null elements.

3.3.2. Constraints

Similarly to the implementation of the cost function, the equalities and inequalities matrices and vectors for the equality and inequalities constrains of (24) are defined. These matrices have $4N + 1$ columns. From (24) the inequalities matrix contain a total of $7N + 6$ rows, 3 of them from (24.a), 3 from (24.b), N from (24.c) and N from each inequality constraint of (24.e-j). Thus the number of elements of the inequality vector is also $7N + 6$. The equality matrix contains only one row due to (24.d) and the equality vector contains only one element.

Each row of inequalities in the constraints of (24) is algebraically equivalent to

$$s_{i,1}b_1 + s_{i,2}b_2 + \dots + s_{i,4N+1}b_{4N+1} \le d_i, \tag{32}$$

where $s_{i,j}$ is the element in row i and column j of the inequalities matrix. The elements of the inequalities vector are d_i. The first 6 values of d_i are e_R, e_A, e_C, $-e_R$, $-e_A$ and $-e_C$, the following $5N$ values are 0 and the final $2N$ values are M. Similarly, the value of the only element of the equality vector is $\#K$.

In order to solve the problem the constants should be specified. The number of audio files used to model specific noise are set to $\#K = 1$. The minimum squared gain coefficient is set to $w_n^{min} = 1 \times 10^{-3}$, the maximum squared gain coefficient is set to $w_n^{max} = 1$. Then $v_n^{min} = w_n^{min}$ and $v_n^{max} = w_n^{max}$ because of (18). Finally the value of the large constant used in (19) was set to $M = 10$.

3.4. Insertion instants and automated audio mixing

For files that contain only one single event or a group of events the insertion instants $t_{i,q}$ are randomly chosen, with uniform probability, from the samples of the output file. This procedure simulates a Poisson distribution of the time intervals.

For files that contain large stationary noises the insertion instants are calculated as

$$t_{i,q} = q \times T_i + T_s \tag{33}$$

for $q \in \{0, 1, 2, \cdots, T_\text{o} \setminus T_i\}$ (where $\setminus$ is the integer division), T_s is the audio sampling period, T_i is the duration of selected file x_i and T_o is the duration specified for output file x_o.

An example of the mixing process has been shown in Figure 1. The first 4 rows contain the audio signals x_i with $i \in I$ affected by the gains g_i and temporally placed at $t_{i,q}$. The last row contains the final mix x_o. The files x_{56} and x_{83} are both long stationary noises. File x_{56} is repeated 3 times and the signal exceeding $T_\text{o} = 60$ s is discarded. File x_{83} is inserted only once because $T_{83} > T_\text{o}$ and the signal exceeding $T_\text{o} = 60$ s is discarded.

We use a set of three rules to avoid the loss of relevant parts of input files. Files in which relevant parts take place when the maximum sound level occurs (e.g. car pass-bys) are inserted matching the sample where the maximum level occurs with the corresponding random insertion instant. Files that contain large stationary noises are inserted by matching the first sample of the input file with each insertion instant calculated following (33). The remaining files are inserted matching the sample in the middle of the input file with the corresponding random insertion instant. The matching sample of each audio file is available from the meta-data file. The interactive operations (iii) include identifying the insertion rule, and the automated operations include determining the matching sample following the corresponding rule.

4. Results

We composed a set of stimuli using the method proposed in Sec 2 in order to evaluate if measured values are close to the values specified for each parameter. In order to compose the stimuli, we used the sound database reported in Section 3.1 and the configuration in Section 3.2 as well as other implementations details reported in Section 3.

The specified values for the parameters of the stimuli correspond to a full factorial experimental design with 4 factors. The set includes all the possible combinations for varying $L_\text{C-A}$ at 5 levels, $L_\text{A-R}$ at 5 levels, $L_\text{A,eq}$ at 2 levels and T_o at 2 levels as shown in Table I. Thus, a set $O = \{x_{\text{o},1}, \ldots, x_{\text{o},n_\text{o}}, \ldots, x_{\text{o},N_\text{o}}\}$ of $N_\text{o} = 5{\times}5{\times}2{\times}2 = 100$ output audio files simulating the total noise was generated. We also generated a set $R = \{x_{\ textr,1}, \ldots, x_{\text{r},N_\text{o}}\}$ of output audio files simulating the corresponding residual noises for each $x_{\text{o},n} \in O$ in order to measure if the specified value of $L_\text{A-R}$ was reached.

Table I. Factors and levels.

Factors	Levels	Minimum	Maximum	Step size
$L_\text{C-A}$	5	2 dB	14 dB	3 dB
$L_\text{A-R}$	5	2 dB	10 dB	2 dB
$L_\text{A,eq}$	2	45 dBA	55 dBA	10 dBA
T_o	2	1 min	3 min	2 min

Table I shows the number of levels at which each factor was varied, the minimum value, the maximum value and the step size for each factor.

In Section 4.1 the results of the mathematical optimization problem are shown and analyzed. In Section 4.2 the measured results are shown as well as several analysis such as kernel density estimation, testing if the measured differences in the factors levels are statistically relevant and correlation analysis. In Sec. 4.3 a subjective test is carried out with a subset of 16 the 100 stimuli in order to assess the realism of the soundscape generator.

4.1. Optimization results

Optimization results are intermediate results of the implementation of this soundscape generator method. These intermediate results show the solution to (24) obtained using CPLEX solver [27]. Sets I and J and gains g_i are estimated from these results. Set I is obtained from its indicator function $\vec{1}_I = (b_{3N+1}, b_{3N+2}, \ldots, b_{4N})$ and set J is obtained from its own indicator function $\vec{1}_J = (b_{2N+1}, b_{2N+2}, \ldots, b_{3N})$. Gains g_i are obtained for $i \in I$ as $g_i = \sqrt{w_i}$, where $w = (b_{N+1}, b_{N+2}, \ldots, b_{N+i}, \ldots, b_{2N})$.

The $N_\text{o} = 100$ instances converged to optimal solutions. A subset of 16 instances reached a feasible optimum by integer tolerance. The integer tolerance was set to 1×10^{-16} in the solver.

The optimal solutions $\hat{\mathbf{u}}$ of (24) matched the required values of $\mathbf{u}$ except for 11 of the 100 solved instances. However, the estimations of $\hat{L}_\text{C-A}$, $\hat{L}_\text{A-R}$ and $\hat{L}_\text{A,eq}$ of these 11 instances are adequate within an error bound of ±0.2 dB (i.e. a value that is below just-noticeable differences for pure tones [28] and quite below the uncertainty of a class 1 sound level meter).

4.2. Stimuli results

The final results are the stimuli as would be measured at a subject's position without the subject. These stimuli are analyzed in terms of the measured values of the $L_\text{A,eq}$, $L_\text{C-A}$ and $L_\text{A-R}$ descriptors of the acoustic signals. The duration of the stimuli are not measured because the possible discrepancies are less than $4 \times 10^{-5}\%$ (i.e. quite below possible perceived duration differences).

The measured values of $L_\text{A,eq}$, $L_\text{C-A}$ and $L_\text{A-R}$ are estimated applying both impulse responses h_o and h_s to both output files x_o and x_r for each of the 100 instances. An audio file example (considering effects of h_s and h_o) for the test of Figure 1 is available in a repository of recordings [29]. It was generated setting $L_\text{A,eq} = 45$ dBA, $L_\text{C-A} = 14$ dB, $L_\text{A-R} = 2$ dB and $T_\text{o} = 1$ min.

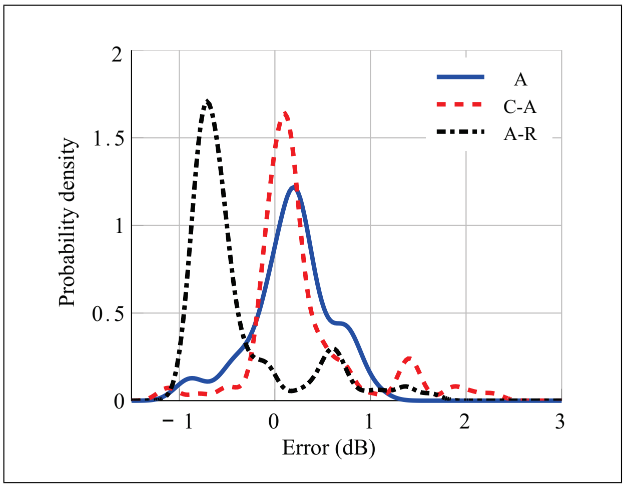


Figure 3. Estimated probability density functions of the errors between specified and measured values. Solid: errors on A-weighted equivalent sound level, Dashed: errors on difference between C-weighted and A-weighted equivalent sound level, Dash-dotted: errors on emergence level.

Let $\check{L}_{\text{A,eq}}$, $\check{L}_{\text{C-A}}$ and $\check{L}_{\text{A-R}}$ be the measured estimations. Thus, the errors between measured and previously specified values for each generated stimulus are $\epsilon_{\text{A}} = \check{L}_{\text{A,eq}} - L_{\text{A,eq}}$, $\epsilon_{\text{C-A}} = \check{L}_{\text{C-A}} - L_{\text{C-A}}$ and $\epsilon_{\text{A-R}} = \check{L}_{\text{A-R}} - L_{\text{A-R}}$. The maximum error is $\epsilon_{\text{C-A},n_{\max}} = 2.3$ dB and the mean error is $\bar{\epsilon} = (\bar{\epsilon}_{\text{A}} + \bar{\epsilon}_{\text{C-A}} + \bar{\epsilon}_{\text{A-R}})/3 = 0.46$.

Figure 3 shows the distribution of errors ϵ_{A}, $\epsilon_{\text{C-A}}$ and $\epsilon_{\text{A-R}}$. The errors are likely associated with violation of the assumption of incoherence between audio files included in each stimulus (See Section 2.2.1), temporal placement process (See Section 2.3) and fractional bands modeling of the scenario configuration (See Section 3.2).

Figure 4 compares the requested values of $L_{\text{A,eq}}$, $L_{\text{C-A}}$, $L_{\text{A-R}}$ and T with the measured values of $\check{L}_{\text{A,eq}}$, $\check{L}_{\text{C-A}}$, and $\check{L}_{\text{A-R}}$. Also a probability density function (scaled for graphical purposes) is shown for each level of each varying factor.

The probability density function (PDF) is estimated using a kernel density estimation technique. Each PDF is placed on the expected level for each corresponding factor, and scaled to reach one half of the corresponding factor step size (See Table I).

Non-parametric methods are preferred because the probability distribution of the measured descriptors for each level is unknown a priori. Figure 4 shows that the distributions are multimodal and difficult to be predicted. For each factor we performed a Wilcoxon rank sum test ($\alpha = 0.05$) by pairs of levels. The null hypothesis, that each pair of tested samples correspond to samples from continuous distributions with equal medians, was rejected for all tested cases.

The correlation of each measured value of $\check{L}_{\text{A,eq}}$, $\check{L}_{\text{C-A}}$, and $\check{L}_{\text{A-R}}$ with each of the specified values of $L_{\text{A,eq}}$, $L_{\text{C-A}}$, $L_{\text{A-R}}$ and T_{o} is shown in Table II.

This correlation matrix shows that the measured values are highly correlated with the corresponding specified values. $\check{L}_{\text{A,eq}}$ dependance of T that shows a small inverse correlation. This correlation is also shown in the first row and last column of Figure 4. Although a correlation coefficient of −0.40 between requested duration T_{o} and A-weighted measured level $\check{L}_{\text{A,eq}}$ could be important for other applications, the $\check{L}_{\text{A,eq}}$ difference is very likely below a just noticeable difference as regards auditory perception. The rest of measured values are not correlate with other controlled factors.

Table II. Correlation coefficients matrix.

	Instance definition			
Measured	$L_{\text{A,eq}}$	$L_{\text{C-A}}$	$L_{\text{A-R}}$	T_{o}
$\check{L}_{\text{A,eq}}$	0.91	0.02	-0.01	-0.40
$\check{L}_{\text{C-A}}$	0.01	1.00	-0.01	-0.03
$\check{L}_{\text{A-R}}$	-0.03	0.02	0.98	0.05

Table III. Simulations used in the realism experiment.

n_{o}	$L_{\text{A,eq}}$ (dB)	$L_{\text{C-A}}$ (dB)	$L_{\text{A-R}}$ (dB)	T_{o} (s)
1	45	2	2	60
2	45	2	10	180
3	45	14	2	180
4	45	14	10	60
5	55	2	2	180
6	55	2	10	60
7	55	14	2	60
8	55	14	10	180

4.3. Realism

We conducted an experiment to assess the realism of the tool using a subset of 8 of the 100 composed stimuli and another set of 4 stimuli directly recorded from a real outdoor environment.

The 8 composed stimuli where chosen to represent the 4 controlled factors in a fractional factorial design (i.e. a 2^{4-1} design) at 2 levels. Table III shows the values of the factors for each stimulus.

The experiment was conducted in the room described in Section 3.2 characterized by its impulse response h_{o}. Figure 2 shows the positions of the source, the participant (P), and the experimenter. The signal played though the loudspeaker is computed as

$$x_{\text{s}}(t) = \left(x_{\text{o}} * h_{\text{s}}\right)(t). \qquad (34)$$

This process simulates that signal x_{o} has passed through a partition characterized by its impulse response h_{s}. This is used for both the simulated and the recorded environmental noises.

The parameters of each stimulus were measured at position P in the room of Figure 2 without subjects. The measured parameters for each stimulus are shown in Table IV.

The 4 stimuli recorded from outdoor environment correspond to different periods of the day in a quiet neighborhood. The main sources are road traffic, dogs, insects and

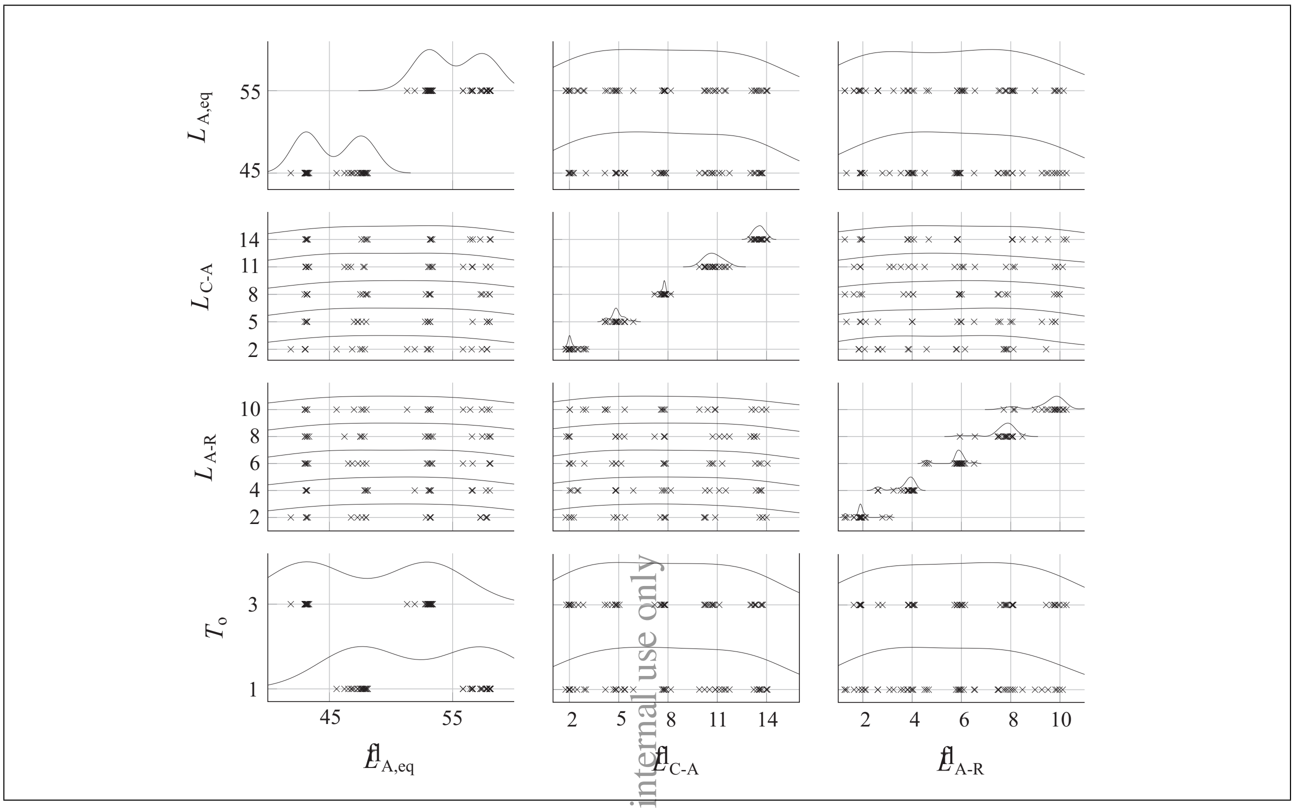


Figure 4. Specified values $L_{\text{A,eq}}$, $L_{\text{C-A}}$, $L_{\text{A-R}}$ and T_{o} relationship with measured descriptors $\check{L}_{\text{A,eq}}$, $\check{L}_{\text{C-A}}$, and $\check{L}_{\text{A-R}}$, all levels are in dB and T_{o} in min. Columns vary for measured values. First column: $\check{L}_{\text{A,eq}}$. Second column: $\check{L}_{\text{C-A}}$. Third column: $\check{L}_{\text{A-R}}$. Rows vary for specified values. First row: $L_{\text{A,eq}}$. Second row: $L_{\text{C-A}}$. Third row: $L_{\text{A-R}}$. Forth row: T_{o}. Cross: a measured value. Line: a Kernel Density Estimation.

birds. Each of the signals were adjusted to $L_{\text{A,eq}}$ of 45 dB or 55 dB. Table V shows the values for $L_{\text{A,eq}}$, $L_{\text{C-A}}$ and T_{o} measured at position P for each stimulus reproduced in the room of Figure 2 after processing according to (34) and without subjects.

The experiment was conducted with 18 participants that auto-reported normal hearing and habits of low exposure to noise. A group of 10 subjects evaluated the simulations and a control group of 8 subjects evaluated the recordings.

The participants of the first group were between 32 and 39 years old with a mean of 36 years and standard deviation of 3 years. The gender distribution of participants was 40% female and 60% male.

The participants of the control group were between 31 and 39 years old with a mean of 35 years and standard deviation of 3 years. The gender distribution of participants in the control group was 38% female and 62% male.

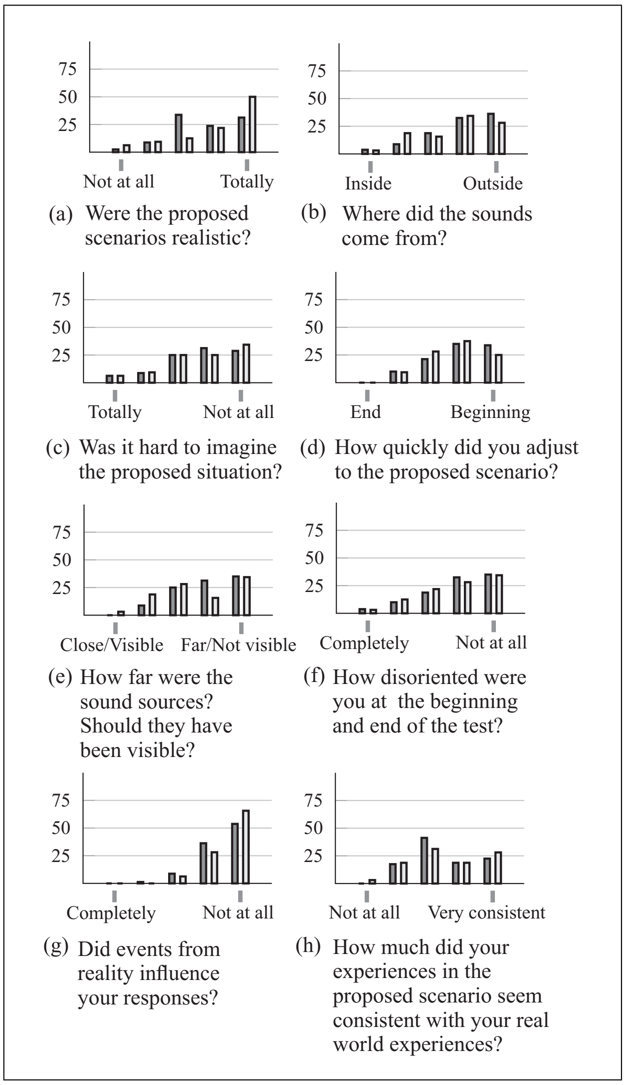
The experiment was performed on one participant at a time. Subjects responded to the questionnaire after the reproduction of each stimulus. Each participant listened the corresponding set of stimuli in a different random order and completed the questionnaire after each stimulus. The participants were instructed to imagine they were in leisure time and in a different scenario each time. The word scenario was defined in the instructions as a situation in a location were they are expected to spend leisure time. The word realistic was instructed to be interpreted as plausible to find in the real world. The control-group subjects received the same instructions and evaluated only the 4 recordings from real environments. The average duration of the subjective test was 25 min for each subject in the first group and 14 min for each subject in the control group.

Table IV. Measured parameters for the simulated stimuli.

n_{o}	$L_{\text{A,eq}}$	$L_{\text{C-A}}$	$L_{\text{A-R}}$
1	47.3	2.0	2.1
2	42.7	2.1	9.4
3	42.8	13.8	1.9
4	47.2	13.7	9.5
5	52.8	2.1	1.8
6	57.0	1.8	11.1
7	56.8	14.0	1.3
8	52.8	13.1	10.1

Table V. Measured parameters for the recordings used with the control group.

$n_{\text{o,c}}$	$L_{\text{A,eq}}$ (dB)	$L_{\text{C-A}}$ (dB)	T_{o} (s)
1	45.0	10.3	60
2	55.0	11.9	180
3	45.0	14.5	180
4	55.0	15.2	60



(a) Were the proposed scenarios realistic?

(b) Where did the sounds come from?

(c) Was it hard to imagine the proposed situation?

(d) How quickly did you adjust to the proposed scenario?

(e) How far were the sound sources? Should they have been visible?

(f) How disoriented were you at the beginning and end of the test?

(g) Did events from reality influence your responses?

(h) How much did your experiences in the proposed scenario seem consistent with your real world experiences?

Figure 5. Responses to questionnaire. Percentage of responses in that point. Dark gray bars: simulated stimuli. Light gray bars: control group.

The questionnaire used in this pilot study of realism is intended to include the actual physical issues and the virtual environment as part of the proposed scenario. The questionnaire is inspired in presence questionnaires used exclusively for virtual environments [2, 10]. The questionnaire was conducted in Spanish. The adaptation of the questionnaire includes translation, selection of the proper questions (i.e. the ones related to sound aspects consistent with this experiment) and the definition of the scale for responses. Responses were collected in 5-point scales similar to the ones reported by Finney and Janer [2].

The questionnaire included the 8 questions listed below with the tags for the first and the last segment in parenthesis.

a) Were the proposed scenarios realistic? (No, not at all - Yes, totally)

b) Where did the sounds came from? (In the room - Out of the room)

Table VI. Questionnaire results: means and standard deviation.

	Simulations		Control	
Question	Mean	Std. dev.	Mean	Std. dev.
a	0.68	0.27	0.75	0.32
b	0.72	0.28	0.66	0.30
c	0.67	0.29	0.68	0.31
d	0.73	0.24	0.70	0.24
e	0.73	0.24	0.65	0.31
f	0.71	0.28	0.70	0.29
g	0.86	0.18	0.90	0.15
h	0.62	0.26	0.63	0.30

Table VII. Questionnaire results: Two-Sample Welch's t-test.

Question	p-Value
a	0.29
b	0.35
c	0.86
d	0.47
e	0.18
f	0.78
g	0.21
h	0.88

c) Was it hard to imagine the proposed situation? (Yes, totally - No, not at all)

d) How quickly did you adjust to the proposed scenario? (Not even when the test ended - At the beginning of the test)

e) How far were the sound sources? Should they have been visible? (Close/Visible - Far/Not visible)

f) How disoriented were you at the beginning and end of the test? (Completely - Not at all)

g) Did the sounds from reality influenced your responses? (Completely - Not at all)

h) How much did your experiences in the proposed scenario seem consistent with your real world experiences? (Not at all consistent - Very consistent)

Figure 5 shows the answers to the questionnaire of realism. For each question, each sub-figure 5.a to 5.h shows the distribution of responses for the group that evaluated the simulations and for the control group.

In order to analyze results, the 5 points scale is represented in a scale from 0 to 1 with steps of 0.25. The mean values and standard deviation for each question are shown in Table VI for the group that evaluated the simulated stimuli as well as for the control group that evaluated the recorded stimuli.

A 2-sample Welch's t-test was performed for each question in order to compare if the mean responses for the simulated stimuli were different respect the stimuli for the control group. The evidence for each question is not sufficient to reject the null hypothesis that the means of both samples are equal ($\alpha = 0.05$). Table VII shows the p-value for the 2-sample Welch's t-test for each question.

The major part of responses to question *a* about realism are distributed from 0.50 to 1.00 in the normalized scale. The unity indicates the proposed scenario is totally realistic.

Responses to question *b* indicate that the major part of the subjects perceived the sounds as generated outside. The condition simulated in (34) is consistent with these responses.

The major part of the responses to question *c* range from 0.50 to 0.75. These results indicate the participants found some difficultly to imagine the proposed situation.

The major part of the responses to question *d* range from 0.75 to 1. No participant responded in 0 point. These results indicate that the major part of participants adjusted quickly to the proposed scenario.

The distribution and mean value of the responses to question *e* is consistent with that of responses to question *b*. The sound sources were perceived as located distant from the participants.

The major part of the responses to question *f* range from 0.50 to 1. These results indicate that the major part of the participants did not experience an important disorientation caused at the beginning and the end of the stimuli.

The major part of responses to question *g* range from 0.75 to 1. This result indicates that participants reported that events from reality (i.e. sounds of clothes, breathing, etc.) did not influence their responses.

The major part of the responses to question *h* range from 0.50 to 0.75. This result indicates the participants found some differences between the proposed scenario and their experiences in the real world. This result seems consistent with that of question *c*. These results are possibly related to realism of stimuli. However, these results could possibly be related also to the physical issues and the situation proposed in the test procedure. Responses about the realism of the scenario itself (question *a*) are favorable but responses on the experience (question *h*) or the situation (question *c*) involving the scenario are less favorable.

5. Conclusions

We developed a method to compose stimuli for future assessment of effects of sound emergence level and low frequency content on annoyance due to environmental noise. The stimuli are composed controlling standardized factors that are frequently used in noise regulations. These factors are the A-weighted equivalent sound level, the emergence level and the difference between C-weighted to A-weighted sound levels. The method solves the problem of controlling the output parameters of a mix of a subset of audio files from a large set of recordings from real sound sources. The control is achieved by determining the parameters of that mix. The mixing-process parameters such as gain and insertion instants applied to each audio file are found solving an optimization problem and using a temporal placement strategy. The whole method could be automated except for a set of interactive operations required to import a new file to the audio files database.

The optimization problem is based on weighted sound exposure combination with several constraints. These constraints are related to realism and to the definition of the residual noise, the specific noise and the total noise.

We implemented the method using a sound database of $N = 97$ audio files and a virtual scenario involving a real room response. A set of $N_\text{o} = 100$ stimuli was composed in order to test the proposed method.

The discrepancy of the numeric solution to the optimization problem is within a bound of ±0.2dB. The measured values for the controlled factors show a mean discrepancy of 0.46 dB and a maximum of 2.3 dB. The mean error could possibly be below just noticeable differences because they may be greater for this kind of stimuli than for pure tones [30] as suggested in [31]. The maximum error can also be close to that limit [31].

A subset of 8 stimuli was evaluated by 10 participants. The main results of subjective evaluation on realism are favorable. The participants found some differences between the proposed scenario and the situations of the real world. These differences could be related to the procedure itself.

The limitations of the proposed method are related to the conditions of the validation experiments, the size of the database and other aspects related to the implementation and context of experiments. If the specified parameters of the stimuli were greater or smaller than the ones used in the validation experiments, the errors would increase unless the database were enlarged including input files with parameters similar to the ones specified for the mix. Subjective results could be improved implementing the method with more realistic auralization systems. The validity of the tool can improve with a more detailed subjective evaluation regarding realism of each stimulus and each sound source listened in each stimulus.

Future work could include potential annoyance assessment of a subset of the stimuli generated in this work or a new experimental design with stimuli generated using the method we propose in this work. An implementation with a larger audio database might possibly reduce the correlation of A-weighted equivalent sound level with duration of the stimulus. However the correlation is small and the effect of stimulus duration on its A-weighted equivalent sound level could be below perceived differences. There is a comprehensive set of attributes in English for spatial audio and a sound quality in general [32]. Future work could include the validation for a translated version or the development of a similar set in Spanish.

Acknowledgement

The authors acknowledge the support from the Consejo Nacional de Investigaciones Científicas y Técnicas (National Scientific and Technical Research Council from Argentina), the Universidad Nacional de Rosario (National University of Rosario from Argentina) and the Universidad Nacional de San Juan (National University of San Juan from Argentina). We are very thankful for the help with language offered by Marcos Toibero.

References

[1] M. Rossignol, G. Lafay, M. Lagrange, N. Misdariis: SimScene: a web-based acoustic scenes simulator. In 1st Web Audio Conference (WAC), 2015.

[2] N. Finney, J. Janer: Soundscape generation for virtual environments using community-provided audio databases. In W3C Workshop: Augmented Reality on the Web, 2010.

[3] A. Valle, V. Lombardo, M. Schirosa: Simulating the soundscape through an analysis/resynthesis methodology. Auditory Display: 6th International Symposium, CMMR/ICAD 2009, Copenhagen, Springer Berlin Heidelberg (2010), 330–357.

[4] S. Innami, H. Kasai: Super-Realistic Environmental Sound Synthesizer for Location-based Sound Search System. IEEE Transactions on Consumer Electronics **57**(4) (2011) 1891–1898.

[5] A. Misra, G. Wang, P. Cook: Musical tapestry: Recomposing natural sounds. Journal of New Music Research, 36(4) (2007) 241–250.

[6] M. Thorogood, P. Pasquier: Computationally created soundscapes with audio metaphor. In Proceedings of the Fourth International Conference on Computational Creativity, (2011) 1–7.

[7] E. Accolti, F. Miyara: Method for generating realistic sound stimuli with given characteristics by controlled combination of audio recordings. J. Acoust. Soc. Am. **137**(1) (2015) EL85–EL90.

[8] M. Alayrac, C. Marquis-Favre, S. Viollon: Total annoyance from an industrial noise source with a main spectral component combined with a background noise. J. Acoust. Soc. Am. **130**(1) (2011) 189–199.

[9] J. Kim, C. Lim, J. Hong, S. Lee: Noise-induced annoyance from transportation noise: Short-term responses to a single noise source in a laboratory. J. Acoust. Soc. Am. **127**(2) (2010) 804–814.

[10] M. J. Singer, B. G. Witmer: Measuring presence in virtual environments: A presence questionnaire. PRESENCE, **7** (1998) 225–240.

[11] ISO 1996–1:2016. Acoustics – Description, measurement and assessment of environmental noise - Part 1: Basic quantities and assessment procedures. International Organization for Standardization, Geneva, 2016.

[12] ISO 1996–2:2007. Acoustics – Description, measurement and assessment of environmental noise - Part 2: Determination of environmental noise levels. International Organization for Standardization, Geneva, 2007.

[13] BS 4142:2014 – Method for Rating Industrial Noise Affecting Mixed Residential and Industrial Areas. British Standards Institution, London, 2014.

[14] IRAM 4062:2016 – Acústica – Ruidos molestos al vecindario. Método de medición y clasificación [Acoustics - Community noise annoyance. Measurement and classification method]. Instituto Argentino de Normalización y Certificación IRAM, Buenos Aires, 2016.

[15] A. Moorhouse, D. Waddington, M. Adams: Proposed criteria for the assessment of low frequency noise disturbance. DEFRA NANR45, Project report, Department for Environment, Food and Rural Affairs 2005.

[16] ANSI/ASA S12.9-2005/Part 4 (R2015) Quantities and procedures for description and measurement of environmental sound – Part 4: Noise assessment and prediction of long term community response, New York, 2015.

[17] A. Kjellberg, M. Tesarz, K. Holmberg, U. Landström: Evaluation of frequency-weighted sound level measurements for prediction of low-frequency noise annoyance. Environ. Int. **23**(4) (1997) 519–527.

[18] H. G. Leventhall: Low frequency noise and annoyance. Noise and Health. **6**(23) (2004) 59–72.

[19] Banco de imágenes y sonidos (Bank of images and sounds). Instituto de Tecnologías Educativas y de Formación del Profesorado. (Institute of Educational Technologies). Ministerio de Educación, Cultura y Deporte (Ministry of Education, Culture and Sports), Spain: Maquina de coser 1 http://mediateca.educa.madrid.org/audio/gzxi8xqcds3lvpys Uploaded: June 12, 2007. Last viewed: November 29, 2016.

[20] J. Sardin: Sound generator 1 kw www.bigsoundbank.com/sound-0115-generator-1-kw.html Uploaded: September 9, 2006. Last viewed: November 29, 2016.

[21] Banco de imágenes y sonidos (Bank of images and sounds). Instituto de Tecnologías Educativas y de Formación del Profesorado. (Institute of Educational Technologies). Ministerio de Educación, Cultura y Deporte (Ministry of Education, Culture and Sports), Spain: Ambiente de viento en la montaña http://mediateca.educa.madrid.org/audio/c6s9q7lmrcrpa6zt Uploaded: June 12, 2007. Last viewed: November 29, 2016.

[22] Banco de imágenes y sonidos (Bank of images and sounds). Instituto de Tecnologías Educativas y de Formación del Profesorado. (Institute of Educational Technologies). Ministerio de Educación, Cultura y Deporte (Ministry of Education, Culture and Sports), Spain: Ambiente de mar con olas 2 http://mediateca.educa.madrid.org/audio/bwnyi3jn5b3ic8w2 Uploaded: June 12, 2007. Last viewed: November 29, 2016.

[23] L. A. Wolsey: Integer Programming. Wiley, 1998.

[24] Banco de imágenes y sonidos (Bank of images and sounds). Instituto de Tecnologías Educativas y de Formación del Profesorado. (Institute of Educational Technologies). Ministerio de Educación, Cultura y Deporte (Ministry of Education, Culture and Sports), Spain. http://recursostic.educacion.es/bancoimagenes/web/. Last viewed: November 29, 2016.

[25] J. Sardin: “Big Soundbank”www.bigsoundbank.com. Last viewed: November 29, 2016.

[26] ISO 18233:2006. Acoustics – Application of new measurement methods in building and room acoustics. International Organization for Standardization, Geneva, 2006.

[27] IBM ILOG CPLEX Optimizer www-01.ibm.com/software/integration/optimization/cp lex-optimizer/ 2011.

[28] J. A. Stillman, J. J. Zwislocki, M. Zhang, L. K. Cefaratti: Intensity just-noticeable differences at equal-loudness levels in normal and pathological ears. J. Acoust. Soc. Am. **93** (1993) 425–434.

[29] E. Accolti: Symtes009S RF www.soundcloud.com/ernesto-accolti/symtes009s-rf. Uploaded: December 14, 2016. Last viewed: December 14, 2016.

[30] H. Fastl, E. Zwicker: Psychoacoustics: Facts and Models. Springer, Berlin, 2007.

[31] F. Fahy, J. Walker (Editors): Fundamentals of Noise and Vibration. E & FN Spon, London, 1998.

[32] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, S. Weinzierl: A Spatial Audio Quality Inventory (SAQI). Acta Acustica United with Acustica **100**(5) (2014) 984–994.