Method for generating realistic sound stimuli with given characteristics by controlled combination of audio recordings

Ernesto Accolti[a] **and Federico Miyara**
Laboratorio de Acústica y Electroacústica, Escuela de Ingeniería Electrónica, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario, Riobamba 245 bis, Rosario, Santa Fe, Argentina
eaccolti@fceia.unr.edu.ar, fmiyara@fceia.unr.edu.ar

Abstract: Audio recordings are often used to improve ecological validity of stimuli for laboratory research on effects of noise. In this paper a method is proposed for composing realistic environmental sound stimuli with (1) specified overall spectrum and (2) specified statistical distribution of sound event durations and semantic categories. The combination is addressed as a mixed integer linear programming problem. Objective measurements, for eight stimuli and a moderate-size database, validate the method. The mean error in octave bands exposure level is 2.6 dB and the statistical distribution of sound event durations and semantic categories is perfectly matched.

[SS]
Date Received: June 27, 2014 **Date Accepted:** December 1, 2014

1. Introduction

A great deal of laboratory research on the effects of noise on humans has been carried out using synthesized sounds, recorded material, or a combination of both. The closer the stimuli are to reality (e.g., recorded material), the greater the ecological validity at the expense of greater difficulty controlling the stimuli parameters. We propose a method to control how recordings are mixed to generate stimuli whose parameters were previously set.

Current studies on the effects of noise are conducted using aural stimuli with controlled parameters such as the spectrum, the sound events duration, occurrence frequency of events, and semantic category of events. For instance, Moorhouse *et al.*[1] assess the influence of the spectrum on the threshold of acceptability of low frequency noise, Vos and Houben[2] assess the influence of the occurrence frequency of impulse events (i.e., short duration) on awakening probability, and Hong and Jeon[3] assess the influence on preference and environmental quality of the semantic category of events conforming the soundscape.

In previous studies, each stimulus is composed starting from small-size audio databases and do not automatically control several parameters together but only groups of a few parameters. For instance, Kim *et al.*[4] control only the broadband level of a single noise event, Vos and Houben[2] generate stimuli by controlling the broadband level and the occurrence frequency of events from a single noise source with only two levels (i.e., one or four repetitions), Alayrac *et al.*[5] mix only two sound files per each stimulus controlling the emergence level, and Hong and Jeon[3] mix only three or less audio files controlling only the semantic category of recordings included in the mix.

We propose a novel method for mixing recordings to compose realistic environmental sound stimuli, controlling the overall band spectrum and the frequency of

[a]Author to whom correspondence should be addressed.

sound events occurrence depending on their semantic category, their duration, or both. The problem is formulated and solved using available optimization techniques[6,7] with integer constraints. The method is applied to recorded material achieving results with acceptable tolerance.

2. Proposed method

Let $W = \{w_1, w_2, \ldots, w_N\}$ be a set of N given audio files, each of them of duration T_n. The problem is to find an index subset $I \subseteq \{1, \ldots, N\}$ and, for each $i \in I$, a gain $g_i \in \mathbb{R}^+$ and a number of repetitions $y_i \in \mathbb{N}$ such that, when the files w_i are mixed together into an output file w_o, we get a realistic environmental noise whose characteristics have been previously required. The audio mix consists of summing the signal of each audio file w_i after applying a gain g_i and repeating it y_i times. Each repetition starts at different times according to one of the three strategies introduced below in Sec. 2.3.

Let $m \in \{1, \ldots, M\}$ denote the mth frequency band and let the sound exposure[8] be computed on an integration period T_u corresponding to the expected duration of the output audio file w_o. Then $\vec{e}_u = (e_{u,1}, \ldots, e_{u,M})^T$ is the required exposure spectrum. Consider that each audio file w_i has a number of sound events of different durations but in the same semantic category and not occurring simultaneously. We shall classify duration into P duration bins, and let $\vec{d}_u = (d_{u,1}, \ldots, d_{u,P})^T$ be the required duration histogram of those events in all the files w_i that will be mixed into w_o. Let S be the total number of semantic categories assigned to the sound files in W. Thus, $\vec{c}_u = (c_{u,1}, \ldots, c_{u,S})^T$ is the required histogram of semantic categories of those events in all the files w_i that will be mixed into w_o.

Some constraints must be introduced to define the whole problem. A first constraint is an upper bond for y_n in order to avoid the unrealistic sensation of perceiving the same sound event many times instead of different sounds of the same kind. A second constraint is an upper bound for g_i to avoid excessive amplification that could be perceived as a source placed closer than usual. A lower bound for g_i should be set to prevent sound events from being masked by other sound events in the file w_o. In order to simplify the problem, we shall consider that g_i and y_i can be set simultaneously to zero in order to exclude a file from the selected subset. Thus, the problem reduces to find the set of gains $g_n \geq 0$ and repetitions $y_n \geq 0$ with $n \in \{1, \ldots, N\}$.

We address the problem using a mixed integer linear programming (MILP) formulation.[6] The objective variables are the number of repetitions $y_n \in \mathbb{N}_0$ and the squared gain coefficients $x_n \in \mathbb{R}_{\geq 0}$, defined as $x_n = g_n^2$ for simplicity, because the sound exposure is proportional to the squared sound pressure. Each file w_i contributes y_i times its sound exposure to the file w_o. This proportionality is accounted for by applying, before mixing, a modified gain $g_i = \sqrt{x_i/y_i}$ instead of the corresponding $\sqrt{x_i}$ in order to avoid nonlinearities in the parameters of the mathematical formulation.

2.1 Linear combination of parameters

Let $d_{p,n}$ be the number of sound events whose duration falls in the time interval corresponding to bin p for each file w_n. Then, the number of sound events in the duration bin p present in the output file w_o can be computed as

$$d_{o,p} = y_1 d_{p,1} + y_2 d_{p,2} + \cdots + y_n d_{p,n} + \cdots + y_N d_{p,N}. \quad (1)$$

Let $c_{s,n}$ be the number of sound events of the file w_n belonging to the semantic category s. Then, the number of sound events belonging to the s category and present in the output file w_o can be computed as

$$c_{o,s} = y_1 c_{s,1} + y_2 c_{s,2} + \cdots + y_n c_{s,n} + \cdots + y_N c_{s,N}. \quad (2)$$

Let $e_{m,n}$ be the mth band sound exposure, calculated over an integration period T_n, for each file w_n. Then, assuming incoherence between w_i files, the exposure in band m of the output file w_o can be computed as

$$e_{o,m} = x_1 e_{m,1} + x_2 e_{m,2} + \cdots + x_n e_{m,n} + \cdots + x_N e_{m,N}. \tag{3}$$

2.2 Optimization formulation

Let the objective variable be a vector $\vec{b} = (x_1, \ldots, x_N, y_1, \ldots, y_N)^T$. Let the required parameters be a vector $\vec{u} = (e_{u,1}, \ldots, e_{u,M}, d_{u,1}, \ldots, d_{u,P}, c_{u,1}, \ldots, c_{u,S})^T$. Let the actual parameters of file w_o be a vector $\vec{o} = (e_{o,1}, \ldots, e_{o,M}, d_{o,1}, \ldots, d_{o,P}, c_{o,1}, \ldots, c_{o,S})^T$ which is expected to reach $\vec{u}$.

Let $E \in \mathbb{R}_{\geq 0}^{M \times N}$ be a matrix with entries $e_{m,n}$, $D \in \mathbb{N}_0^{P \times N}$ be a matrix with entries $d_{p,n}$, and $C \in \mathbb{N}_0^{S \times N}$ be a matrix with entries $c_{s,n}$. Thus we define matrix $A \in \mathbb{R}_{\geq 0}^{(M+P+S) \times 2N}$ as

$$A = \begin{bmatrix} E & \mathbf{0}_{(M,N)} \\ \mathbf{0}_{(P,N)} & D \\ \mathbf{0}_{(S,N)} & C \end{bmatrix}, \tag{4}$$

where $\mathbf{0}_{(h,k)}$ are $h \times k$ sized matrices of null elements.

We minimize $|e_{u,1} - e_{o,1}|, \ldots, |e_{u,M} - e_{o,M}|, |d_{u,1} - d_{o,1}|, \ldots, |d_{u,P} - d_{o,P}|, |c_{u,1} - c_{o,1}|, \ldots,$ and $|c_{u,S} - c_{o,S}|$ by minimizing the cost function $f_s = \|A \times \vec{b} - \vec{u}\|_\infty$. The ℓ_∞ norm is proposed because small differences distributed in the parameters could be closer to psychoacoustic just-noticeable differences.[9]

Since the cost function f_s is non-linear, a linear ℓ_∞ approximation[6,7] is applied below. Let $v \in \mathbb{R}_{\geq 0}^{+}$ and $\vec{b}$ be the objective variables and let $f_l(\vec{b}, v) = v$ be the new linear cost function. Then the ℓ_∞ approximation is formulated as the following MILP problem:

$$\begin{aligned}
&\text{minimize} && v \\
&\text{subject to} && A \times \vec{b} - \vec{u} \leq v\mathbf{1}, && \text{(5a)} \\
& && -(A \times \vec{b} - \vec{u}) \leq v\mathbf{1}, && \text{(5b)} \\
& && y_n \leq y_n^{\max},\ n = 1, 2, \ldots, N, && \text{(5c)} \\
& && x_n \leq x_n^{\max} \times y_n,\ n = 1, 2, \ldots, N, && \text{(5d)} \\
& && -x_n \leq -x_n^{\min} \times y_n,\ n = 1, 2, \ldots, N, && \text{(5e)} \\
& && x_n, v \in \mathbb{R}_{\geq 0} \text{ and } y_n \in \mathbb{N}_0,\ n = 1, 2, \ldots, N, && \text{(5f)}
\end{aligned}$$

where $\mathbf{1}$ is a vector of dimension $M + P + S$ with all elements equal to unity.

The ℓ_∞ norm approximation is described by Eqs. (5a) and (5b) and the cost function. Equation (5c) sets the upper bound $y_n^{\max}$ for y_n. Equations (5d) and (5e) set an upper bound $x_n^{\max}$ and a lower bound $x_n^{\min}$, respectively, for x_n. Notice that x_n can be set to 0 if and only if y_n is set to 0 too, within x_n bounds.

2.3 Insertion instants and audio mixing

Two strategies are used to achieve realistic temporal distribution of the sound events. When composing the set W, each w_i file is edited to (1) contain only one event in case it happens to sound realistic when temporally distributed following a Poisson distribution or (2) to contain a group of similar events when the temporal distribution is unknown. The two classes of files are distributed following a Poisson distribution when mixing (notice that files containing a group of events will internally follow the recorded distribution, not necessarily a Poisson distribution).

A third strategy is used for long stationary noises. These files are prepared to be looped using known audio editing techniques to avoid clicks (i.e., phase breaks). Before defining the matrix E, the exposure of these long events is scaled to the duration T_l of the output file instead of T_n (using weight T_l/T_n for all m). Finally, when mixing, these files are automatically looped, cut, or both, depending on T_l and T_n.

3. Objective validation of the method

3.1 Setup

Eight files containing environmental noises were composed by solving the formulation given in Eq. (5). The MILP package of CPLEX SOLVER[10] was used. CPLEX is a state-of-the-art mathematical programming tool designed for the resolution of MILPs (among others). Each output audio file is $T_l = 120$ s long and has a sample rate of 44 100 Hz. The vector of required parameters $\vec{u}$ for each test is shown in each column of Table 1. The first $M = 7$ rows, which correspond to the required exposure $\vec{e}_u$ in 1/1 octave bands centered from $f_{c,1} = 63$ Hz to $f_{c,7} = 4$ kHz, are shown in terms of exposure level[8] with integration period of 120 s. The next $P = 5$ rows correspond to $\vec{d}_u$ elements which contain the required number of events in each duration bin. The five duration bins are the following intervals in seconds]0.0; 1.0],]1.0; 4.9],]4.9; 12.5],]12.5; 23.5], and $]23.5; \infty]$. The last row indicates that sound events of category "indoors" should be excluded (i.e., only constraint for s corresponding to category "indoors" was set). These required semantic categories $\vec{c}_u$ specify that the files w_n containing sound events usually found indoors should be excluded in I, and in turn excluded in w_o.

A set W of $N = 100$ audio files was collected and each w_n was edited following the strategies for temporal distribution (see Sec. 2.3). The duration of every event in each file w_n was computed in accordance with ISO 1996-1 (Ref. 8) and binned to define the nth entry of matrix D. The number of sound events usually found indoors for each file w_n was aurally detected in order to define the nth entry of matrix C. The 1/1 octave band exposure spectrum of each w_n file was computed to define the nth entry of matrix E.

The maximum number of repetitions $y_n^{\max}$ was set to 1 for long stationary noises and 3 for the remaining files. The maximum squared gain coefficients $x_n^{\max}$ for each $n \in \{1, \ldots, N\}$, and for each test, was set to ensure that the maximum level of each file w_i is 7 dB above the average sound level $[10 \log(\vec{e}_u / T_u)]$ for the band in which this difference is the greatest. The minimum $x_n^{\min}$ was set to ensure that the maximum level of each file w_i is 10 dB below the average sound level for the band in which these values are the closest.

The playback system, consisting of a sound interface connected to an active loudspeaker hidden behind a glass window with a wooden shutter, was characterized by a measured calibration constant k and an inverse filter for loudspeaker-subject path, including the real room response. The inverse filter and the calibration constant k were applied to each entry of matrix E before defining matrix A in Eq. (4).

Table 1. Required parameters for eight environmental noises.

	test							
Parameters	1	2	3	4	5	6	7	8
$L_{e_{u,1}}$	79.7	79.7	89.7	89.7	84.0	84.0	94.0	94.0
$L_{e_{u,2}}$	74.2	74.2	84.2	84.2	76.5	76.5	86.5	86.5
$L_{e_{u,3}}$	68.7	68.7	78.7	78.7	69.0	69.0	79.0	79.0
$L_{e_{u,4}}$	63.2	63.2	73.2	73.2	61.5	61.5	71.5	71.5
$L_{e_{u,5}}$	57.7	57.7	67.7	67.7	54.0	54.0	64.0	64.0
$L_{e_{u,6}}$	52.2	52.2	62.2	62.2	46.5	46.5	56.5	56.5
$L_{e_{u,7}}$	46.7	46.7	56.7	56.7	39.0	39.0	49.0	49.0
$d_{u,1}$	5	4	8	1	8	1	5	4
$d_{u,2}$	4	5	7	2	7	2	4	5
$d_{u,3}$	3	6	6	3	6	3	3	6
$d_{u,4}$	2	7	5	4	5	4	2	7
$d_{u,5}$	1	8	4	5	4	5	1	8
$c_{u,s_{\text{indoor}}}$	0	0	0	0	0	0	0	0

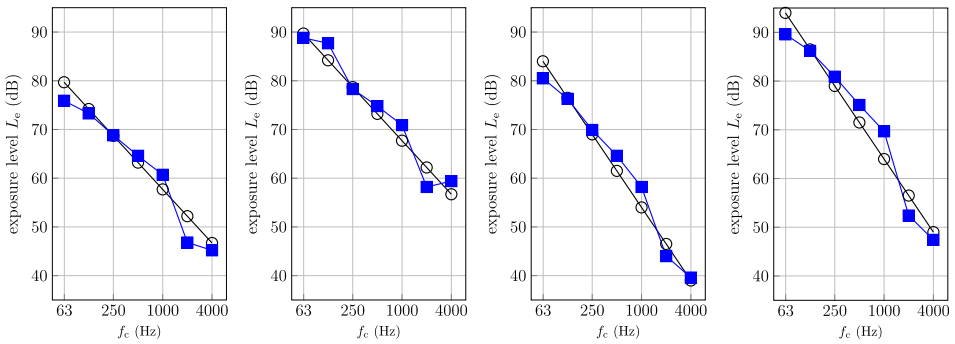

Fig. 1. (Color online) Exposure spectra of tests 1, 3, 5, and 7. Circles: required. Squares: measured.

3.2 Results

The specified number of events in each p-bin of duration and the semantic constraint were perfectly matched for the eight tests. The exposure reached by the output files slightly differs from the required exposure. Figure 1 shows the required exposure (circles) and the exposure measured with a class 1 sound level meter with octave band analyzer (squares). These differences are explained in part by the optimization itself (i.e., the optimum does not necessary match the required parameters), in part because the assumption of incoherence between w_i files is not being fulfilled, and in part because of the frequency resolution of the filter (i.e., 1/1 octave bands) accounting for the playback system. The absolute difference between measured and required exposure in octave bands has a mean value of 2.6 dB, a standard deviation of 2.0 dB and a maximum value of 7.4 dB. The mean value of the cost function across the tests was $\bar{v} = 3.3521$ or $L_{\bar{v}} = 5.3$ dB in order to compare with exposure level differences.

4. Concluding remarks

The novelty of this work is the generalization of the problem of mixing audio files to compose a realistic environmental noise stimulus with previously defined parameters. This generalization is formulated as a MILP problem and several strategies for audio mixing.

The proposed method provides a tool for generating sets of environmental sound stimuli with parameters values set according to the design of an experiment. The method has been implemented in a computer algorithm to automatically generate large sets of stimuli for experiments requiring assessment of human response to several levels of the controlled factors.

Results validate the method from an objective standpoint. Although a 7.4 dB difference in one octave band is noticeable for stationary noises, it could probably be below just-noticeable differences for long-term average spectrum of environmental noise. The difference caused by the optimization itself can be reduced using a larger database or removing constraints (e.g., making frequency bands and duration intervals wider and reducing semantic constraints) depending on each experimental hypothesis.

Acknowledgments

This work was funded by the Argentine Agency of Scientific and Technological Promotion (ANPCyT). The authors especially thank Graciela Nasini, Daniel Severín, and Ernesto Kofman.

References and links

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

[2]J. Vos and M. M. J. Houben, "Enhanced awakening probability of repetitive impulse sounds," J. Acoust. Soc. Am. **134**(3), 2011–2025 (2013).
[3]J. Y. Hong and J. Y. Jeon, "Designing sound and visual components for enhancement of urban soundscapes," J. Acoust. Soc. Am. **134**(3), 2026–2036 (2013).
[4]J. Kim, C. Lim, J. Hong, and 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), 804–814 (2010).
[5]M. Alayrac, C. Marquis-Favre, and 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), 189–199 (2011).
[6]L. A. Wolsey, *Integer Programming* (Wiley, New York, 1998).
[7]S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004).
[8]ISO 1996-1:2003: Acoustics—Description, Measurement and Assessment of Environmental Noise—Part 1: Basic Quantities and Assessment Procedures (International Organization for Standardization, Geneva, 2003).
[9]H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models* (Springer, Berlin, 2005).
[10]IBM ILOG CPLEX Optimizer (ILOG, 2011) [computer program], www-01.ibm.com/software/integration/optimization/cplex-optimizer/ (Last viewed December 9, 2014).