Proceedings of the 2011 International Conference on Auditory Display, somewhere, sometime, 2011

AUDITORY RECOGNITION OF FLOOR SURFACES BY TEMPORAL AND SPECTRAL CUES OF WALKING

Federico Fontana, Fabio Morreale, Tony Regia Corte, Maud Marchal, Anatole Lecuyer

Final manuscript
Access limited to private web site area
federico.fontana@uniud.it,

ABSTRACT

In a multiple choice auditory experimental task, listeners had to discriminate walks over floors made of concrete, wood, gravel, or dried twigs. Sound stimuli were obtained by mixing temporal and spectral signal components, resulting in hybrid formulations of such materials. In this way, we analyzed the saliency of the corresponding cues of time and frequency in the recognition of a specific floor. Results show that listeners differently weigh such cues during recognition, however this tendency is not polarized enough to enable interaction designers to reduce the functionality of a walking sound synthesizer to simple operations made on the temporal or spectral domain depending on the simulated material.

1. INTRODUCTION

Walking sounds play a significant role in everyday listening environments. They in fact convey information at different levels of approaching walkers, and about the surface they traverse. By listening to a temporal sequence of footsteps humans try to identify surface material [1], meanwhile they make inferences on shoe types and walking styles to support decisions on gender as well as about physical, biomechanical, and affective characters of a person [2, 3, 4].

For their importance in the auditory scene, walking sounds have found place in multimodal displays since the early days of interactive simulation. All vintage electronic game players probably remember the iconic use of footstep sounds to render the number and moving speed of the enemies in Space Invaders™, a popular electronic game of the late 70's. Nowadays, with increasing computational and memory resources as well as quality of the pc audio hardware, sound designers can make use of rich collections of accurate walking sounds recorded under different conditions—for instance see the related section in sounddogs.com. Furthermore, techniques exist for the interactive synthesis of footstep sounds [5, **?**, **?**].

Walking sounds lose their meaning if they are displayed inaccurately in an interactive context. When the feedback is predictable with some hundreds of milliseconds, or when a comparable latency of the response to an input is allowed, then prerecorded footstep sounds can be accurately selected from a sample database and even post-processed before reproduction. Conversely, there are situations in which no more than few tens of milliseconds are allowed to an interactive computer system for delivering sounds. In the most compelling case, feedback must be provided under continuously varying control conditions. Consider, for instance, a virtual environment where users can walk across a ground possessing varying degrees of resistance, such as gravel or mud: In this case, the use of physically based or physically informed synthesis models such as those mentioned before becomes unavoidable.

Psychoacoustics has demonstrated that auditory information in sounds is redundant—think, for instance, to the well-known mp3 audio coding. As a consequence, research in interactive synthesis looks for models providing accurate sounds and simple access to their distinctive parameters, meanwhile being parsimonious in terms of needed computational and memory resources. Successful models usually base their performances on applicable experimental results drawn from psychoacoustics. In our specific case, we would be ultimately able to synthesize realistic interactive walking sounds by means of a simple model, whose computation and continuous control is at reach of current consumer hardware. Our basic hypothesis, elaborated in the next section, is that perceived material properties in walking sounds map to a different extent to temporal or spectral auditory cues depending on the nature and, hence, category of the floor material.

2. EXPERIMENTAL HYPOTHESIS

Footstep sounds range across a number of possibilities depending on the shoe type, nature of the floor, and foot action. We restrict the pool of possible interactions to those generated by a male walking with normal style upon a flat floor using leather shoes, hence varying only the floor material. We categorize floors into *solid* or *aggregate*: the former, such as concrete, marble, and wood, are stiff; the latter, such as gravel, dry leaves, and sand, allow relative motion of their constituent units and progressively adapt to the sole profile during the interaction.

Solid materials gives rise to short, repeatable impacts having a definite spectral color. Conversely, aggregate materials elicit sequences of tiny impacts of distinctive temporal density that creates a sort of "crumpling", less resonant sound. We hypothesize that *spectral* cues are more salient in the recognition of solid materials, conversely *temporal* cues are more salient in the recognition of aggregate materials. In particular, we experiment using concrete (C) and wooden (W) floors, representative of solid materials, as well as with gravel (G) and dried twigs (T), representative of aggregate materials. Figure 1 illustrates the hypothesis.

3. METHOD

3.1. Setup

The experiment was set up in the VIPS laboratory of the Dipartimento di Informatica, University of Verona. Subjects were sitting

MATERIAL	PHYSICAL PROPERTIES	ACOUSTIC PROPERTIES
C , W	Solid	Spectral Cues
G , T	Aggregate	Temporal Cues

Figure 1: *Experimental hypothesis. (C: concrete, W: wood, G: gravel, T: twigs.)*

in front of a Mac Pro pc running a Java application communicating (via the *pdj* library) with Pure Data, a free software environment for real time audio synthesis also enabling simple visualizations (through the *gem* library). They listened to the auditory stimuli through a pair of AKG K240 headphones.

3.2. Participants

Thirteen male and three female undergraduate Italian computer science students aged aged 22 to 31 (mean = 24.62, std = 2.55) participated in the experiment. Few of them had some experience in sound processing. All of them reported to usually wear snickers.

At the end of the experiment, every subject completed a subjective questionnaire about the audio stimuli.

3.3. Procedure

One footstep by a normally walking male wearing leather shoes was repeatedly recorded while he stepped over a tray filled with gravel and, then, dried twigs. Recordings were made inside a silent, normally reverberant room using a Zoom H2 digital hand recorder standing 0.5 m far from the tray. For each of the two materials, seven recordings were selected and randomly enqueued to create 12 s long walking sequences, containing 13 footsteps. In addition to these recordings, high quality samples of a male walking on concrete and on parquet wood were downloaded from the commercial database `sounddogs.com`. Using these samples, two further walking sequences were created having the same beat of the footsteps and SPL (??? dB, measured by a Cesva ??? phonometer located in the middle of the two headphone transducers).

Temporal envelopes were extracted from every sequence, by computing the signal

$$e_M[n] = (1 - b[n])|s_M[n]| + b[n]e_M[n-1] \qquad (1)$$

out of the corresponding sequence s_M, $M \in \mathcal{M} = \{C, W, G, T\}$. (Refer to Figure 1 for the meaning of the C, W, G, and T.) As in previous research on synthetic footsteps, the envelope following parameter $b[n]$ was set to 0.8 when $|s_M(n)| > e_M(n-1)$ and to 0.998 otherwise [**?**].

By dividing every sequence s_M by its envelope e_M, we computed signals $u_M = s_M/e_M$ in which the temporal dynamics was removed. In other words, we derived new footstep sequences having stationary amplitude along time. What remained in u_M was a color, that we extracted by computing the coefficients of a 48th-order inverse LPC filter model h_M^{-1} in correspondence of those parts of the signals containing footstep sounds. In the end, for

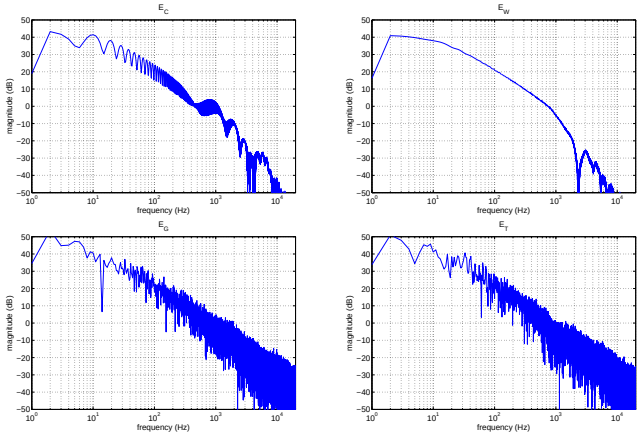

Figure 2: *Magnitude spectra of envelopes E_C, E_W, E_G, and E_T. Respective dc components removed for ease of inspection.*

every material M a highly realistic version $\tilde{s}_M$ of the original sequence s_M could be resynthesized by simply convolving digital white noise n by the "coloring" filter h_M, and then multiplying its output, i.e. the synthesized version $\tilde{u}_M$ of u_M, by the envelope signal e_M:

$$\tilde{s}_M[n] = (n * h_M)[n] \cdot e_M[n] = \tilde{u}_M[n] \cdot e_M[n]. \qquad (2)$$

This technique draws ideas from a family of physically-informed models of walking sounds [**?**, **?**]. In the meantime it provides a simpler, more controlled resynthesis avoiding stochastic generation of patterns as in such models.

Sixteen stimuli were finally created by adding twelve *hybrid* resyntheses to the *native* stimuli $\tilde{s}_C$, $\tilde{s}_W$, $\tilde{s}_G$, and $\tilde{s}_T$. Every hybrid stimulus $\tilde{s}_{M_t,M_f}$, $M_t, M_f \in \mathcal{M}$ was defined as to account for the spectral color of material M_f *and* the temporal envelope of material $M_t \neq M_f$:

$$\tilde{s}_{M_t,M_f}[n] = (n * h_{M_f})[n] \cdot e_{M_t}[n] = \tilde{u}_{M_f}[n] \cdot e_{M_t}[n]. \qquad (3)$$

For each material M_f, we checked that all hybrid temporal manipulations using $M_t \neq M_f$ did not notably alter the spectral information of $\tilde{s}_{M_f}$, and thus its original color. In fact, an inspection of the spectra $E_M(\omega)$ of the various envelopes made by Fourier-transforming e_M, i.e., $E_M(\omega) = \mathcal{F}\{e_M\}(\omega)$, shows that they all have a comparable spectrum. More precisely, all spectra E_C, E_W, E_G, E_T exhibit similar magnitudes, that are shown in Figure 2 after removing the respective dc component for ease of inspection. Hence, spectral differences in $\tilde{S}_{M_t,M_f}(\omega)$ caused by multiplying $\tilde{u}_{M_f}$ by e_{M_t}, that is,

$$\tilde{S}_{M_t,M_f}(\omega) = \mathcal{F}\{\tilde{u}_{M_f} \cdot e_{M_t}\}(\omega) = (\tilde{U}_{M_f} * E_{M_t})(\omega), \qquad (4)$$

are almost the same as those introduced in $\tilde{s}_{M_f}$ by its own envelope e_{M_f}, independently of the material.

Symmetrically, the temporal artifacts which are caused by hybridization can be considered minor. In fact, because of the LPC design methodology, all filters h_C, h_W, h_G, h_T do transform white noise into a stationary signal independently of the material.

3.4. Protocol

Every participant was involved in 192 trials, resulting by randomly playing twelve times each of the sixteen synthetic stimuli. To introduce the auditory stimuli of walking as well as the experimental interface, all subjects were asked to select and play several times each one of the original sequences s_C, s_W, s_G, s_T for some minutes before the experiment.

Figure 3 shows the graphic buttons used in the experiment. At each trial the subject listened to a stimulus. After or during

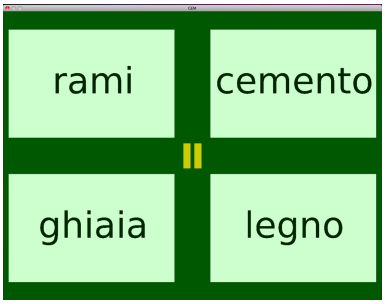

Figure 3: *Graphical interface used in the experiment. (cemento = C, legno = W, ghiaia = G, rami = T.)*

listening, she/he selected the material that was best represented by the corresponding sound learned during the training phase. The selection was made by clicking the related button. As soon as a material was selected the screen froze for a couple of seconds and changed color, to inform subjects of the conclusion of the trial. After this pause, another sound sequence was picked up and a new trial began.

The four buttons randomly switched position at each trial. Subjects could temporarily stop the experiment by clicking the pause icon '||' located in the middle of the screen, whenever they wanted to take a short break among trials. It took approximately ??? minutes for each participant to complete the whole session.

4. RESULTS

For each participant, the percentages of choice for the four materials C, W, G, T were calculated on the 12 trials for the sixteen audio stimuli. A first analysis considered only the participants showing an auditory recognition of the four native stimuli $\tilde{s}_C$, $\tilde{s}_W$, $\tilde{s}_G$, and $\tilde{s}_T$ being significantly higher than random (i.e., 25%). Thus, as the critical value (with $\alpha = 0.05$) of the one-tailed binomial test Bin(12,0.25) is 7 trials (i.e., 58.33%), the participants with an auditory recognition lower than 58.33% for the native stimuli were excluded from the analysis. Using this criterion, 16 participants were considered for the recognition of dried twigs, 15 for gravel, 16 for wood, and 10 for concrete.

The results are presented in Figure 4. In these plots, a bar exhibiting a low percentage means that the correspondingly manipulated information (either temporal or spatial) is important for the recognition of the original material, represented by the leftmost bar in the same plot. The difference from random percentage (25%) was tested using one-proportion (two-tailed) z tests.

In another analysis, we evaluated the auditory recognition of aggregate and solid material categories. Again, this analysis was conducted with the participants exhibiting an auditory recognition significantly higher than random for the two sets of stimuli

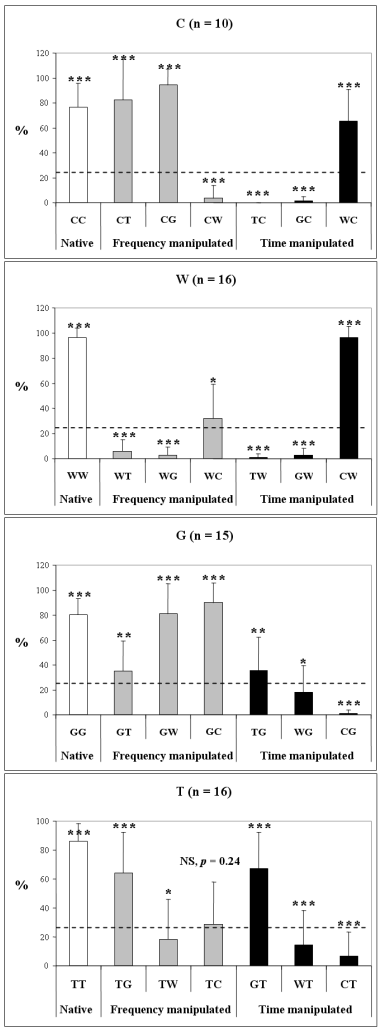

Figure 4: *Mean percentages of choice (bars represent std) for C,W,G,T as a function of the auditory stimulus $\tilde{s}_{M_t,M_f}$. The difference from random choice (line at 25%) was tested using one-proportion (two-tailed) z tests. Note: $*: p < 0.05$, $**: p < 0.01$, $***: p < 0.001$, NS: not significant, n: number of subjects.*

accounting for the respective categories (24 trials for each category). In this case, the critical value (with $\alpha = 0.05$) of the one-proportion (one-tailed) z test is 10 trials, corresponding to 41.67%.

All the participants (n = 16) were selected with this criterion. The results are presented in Figure 5. For the different percentages of choice, the difference relative to random (25%) was tested using one-proportion (two-tailed) z tests. Thus, for the Aggregate category, the percentages of choice in native (82.29%) and frequency manipulated (52.78%) conditions were significantly different from random ($z = 25.98$, $p < 0.001$ and $z = 21.77$, $p < 0.001$, respectively). By contrast, the percentage of choice in the time manipulated condition (23.44%) was not significantly different from random ($z = -1.22$, $p = 0.22$). On the other hand, for the Solid category, the percentages of choice in native (78.39%) and frequency manipulated (35.59%) conditions were significantly different from random ($z = 24.16$, $p < 0.001$ and $z = 8.30$,

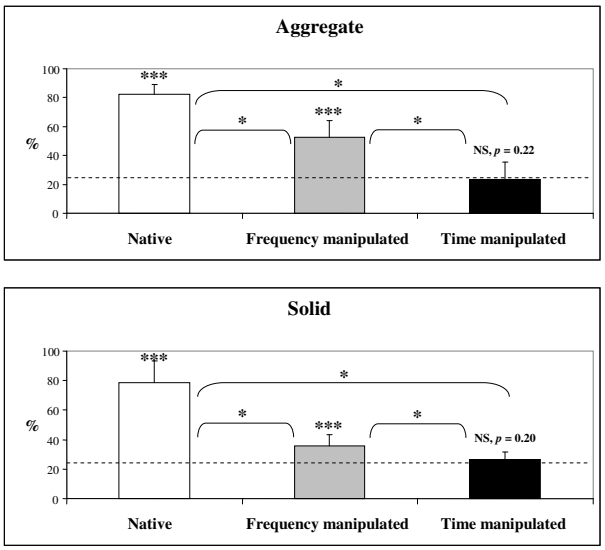

Figure 5: *Mean percentages of choice (bars represent std) for material categories (Aggregate and Solid) as a function of the auditory stimulus $\tilde{s}_{M_t,M_f}$. The difference from random choice (line at 25%) was tested using one-proportion (two-tailed) z tests. The differences between the three audio conditions were tested with two-proportion z tests (two-tailed and Bonferroni-adjusted alpha level with $p = 0.05/3 = 0.0167$). Note: $*: p < 0.05$, $**: p < 0.01$, $***: p < 0.001$, NS: not significant.*

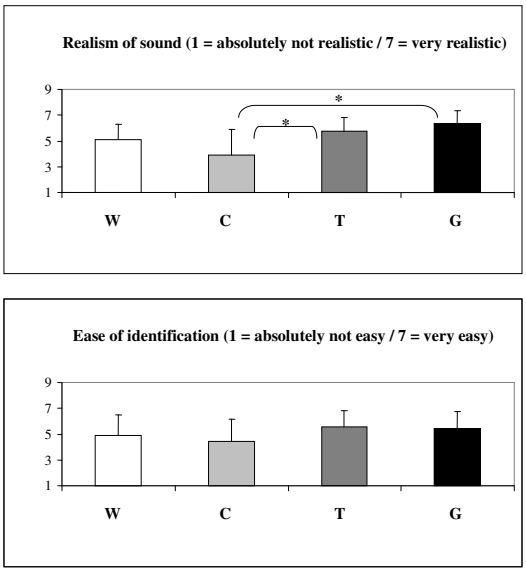

Figure 6: *Mean and standard deviation of subjective ratings about "realism of sound" and "ease of identification" for the four materials C, W, G, T. Differences between materials were tested with two-proportion z tests (two-tailed and Bonferroni-adjusted alpha level with $p = 0.05/6 = 0.0083$). Only the significant differences are presented, $*: p < 0.05$.*

$p < 0.001$, respectively). By contrast, the percentage of choice in the time manipulated condition (26.65%) was not significantly different from random ($z = 1.29$, $p = 0.20$). The differences between the three audio conditions were tested with two-proportion (two-tailed) z tests.

A correction for experiment-wise error was realized by using Bonferroni-adjusted alpha level (p divided by the number of tests). Thus, in order to compare the three audio conditions (native, frequency manipulated, and time manipulated), the alpha level was adjusted to $p = 0.05/3 = 0.0167$. For the Aggregate category, the analysis showed that the native condition was significantly different from the frequency manipulated ($z = 10.23$, $p < 0.05$) and time manipulated ($z = 20.56$, $p < 0.05$) conditions. The difference between frequency manipulated and time manipulated conditions was significantly different ($z = 14.50$, $p < 0.05$) as well. For the Solid category, the analysis indicated that the native condition was significantly different from the frequency manipulated ($z = 14.57$, $p < 0.05$) and time manipulated ($z = 17.96$, $p < 0.05$) conditions. The difference between frequency manipulated and time manipulated conditions was also significantly different ($z = 4.63$, $p < 0.05$).

After the experiment, a questionnaire was proposed in which each participant had to grade from 1 to 7 the four native stimuli according to two subjective criteria: realism, and ease of identification. Figure 6 shows the means and standard deviations of the four native stimuli for each of the subjective criteria. Wilcoxon signed rank (two-tailed) tests with Bonferroni correction showed significant differences only for the realism of sounds: between concrete and dried twigs ($z = -3.28$, $p = 0.001$), and between concrete and gravel ($z = -3.16$, $p = 0.0016$).

5. DISCUSSION

The histograms for concrete and wood in Figure 4 show that subjects tolerate swapping between the temporal features of C and W, conversely the substitution in the same signals with temporal features extracted from aggregate materials (i.e. G and T) harms the recognition. This result is in favor of the initial hypothesis. The effect of spectral manipulations of C and W is more articulate. In this case the hypothesis is essentially confirmed with wood, whose distinctive color cannot be changed using any other spectrum. In parallel, subjects are tolerant to substitutions in C with spectra from aggregate materials. This greater tolerance may be due to the basic lack of distinct character of concrete floors, especially for listeners who usually wear rubber sole shoes such as snickers (indeed the majority of our sample). The same conclusion finds partial confirmation by Figure 6 which, in the limits of the significance of its data, shows greater confidence by subjects in recognizing aggregate materials.

The histograms in Figure 4 regarding gravel and twigs partially support the initial hypothesis. Time swaps between G and T are tolerated to a lesser extent compared to solid floors. Like before, substituting the temporal features of solid materials in an aggregate sound is not tolerated. Spectral substitutions are not as destructive as they were for solid materials, especially in the case of gravel. The worst situation is when the spectrum of W is substituted in T, again probably due to the distinct color that wood resonances bring into the sound.

Figure 5 would further support this discussion. In fact, in spite of the low significance of the data from time manipulations (i.e. black bars), it shows that subjects are primarily sensitive to temporal substitutions between solid and aggregate materials. In parallel, spectral changes are more tolerated during the recognition

of aggregate material compared to solid floors.

6. CONCLUSIONS

The proposed experiment has confirmed that solid and aggregate floor materials exhibit precise temporal features, that cannot be interchanged while designing accurate walking sounds. Within such respective categories, color represents an important cue for the recognition of solid materials, conversely sounds of aggregate materials seems to tolerate larger artefacts in their spectra.

Finally, the proposed resynthesis model could be easily turned into an interactive walking sound synthesizer, by just controlling at runtime (typically by means of force signals acquired from sensing shoes) the reproduction speed and amplitude of temporal envelopes, weighing LPC-filtered noise depending on the simulated material.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme under FET-Open grant agreement 222107 NIW - Natural Interactive Walking.

8. REFERENCES

[1] B. L. Giordano, S. McAdams, Y. Visell, J. R. Cooperstock, H. Yao, and V. Hayward, "Non-visual identification of walking grounds," in *Proc. of Acoustics'08 in J. Acoust. Soc. Am.*, 2008, vol. 123 (5), p. 3412.

[2] X. Li, R. J. Logan, and R. E. Pastore, "Perception of acoustic source characteristics: Walking sounds," *Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 3036–3049, 1991.

[3] Y. Visell, F. Fontana, B.L. Giordano, R. Nordahl, S. Serafin, and R. Bresin, "Sound design and perception in walking interactions," *Int. J. Human-Computer Studies*, , no. 67, pp. 947–959, 2009.

[4] R. Bresin, A. de Witt, S. Papetti, M. Civolani, and F. Fontana, "Expressive sonification of footstep sounds," in *Proc. of the Interaction Sonification workshop (ISon) 2010*, R. Bresin, T. Hermann, and A. Hunt, Eds., KTH, Stockholm, Sweden, Apr. 7 2010.

[5] P. R. Cook, "Modeling Bill's gait: Analysis and parametric synthesis of walking sounds," in *Proc. Audio Engineering Society 22 Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, July 2002, AES.