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

Federico Fontana

University of Udine
Dipartimento di Matematica e Informatica
via delle Scienze 208
33100 Udine, Italy
federico.fontana@uniud.it

Fabio Morreale

University of Trento
Dipartimento di Ingegneria
e Scienza dell'Informazione
via Sommarive 14
38123 Povo (TN), Italy
fabio.morreale@disi.unitn.it

Tony Regia-Corte, Anatole Lécuyer, Maud Marchal

INRIA
Campus Universitaire de Beaulieu
35042 Rennes, France
tony.regia-corte@wanadoo.fr
{Anatole.Lecuyer,Maud.Marchal}@inria.fr

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 computer 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, 6, 7].

Walking sounds lose ecological coherence if they are inaccurately displayed in an interactive context, in terms of audio quality or temporal synchronization. When the feedback message can be completely determined some hundreds of milliseconds before displaying, or when a comparable latency of the response to an input is tolerated, then accurate pre-recorded footstep sounds can be selected from an even huge sample database, and post-processed before reproduction also by devices whose memory and computation power are relatively smaller, such as mobile platforms. Conversely, there are situations in which no more than few tens of milliseconds are allowed to a computer system for delivering interactive sounds. In the most compelling case, feedback must be provided under continuously varying control conditions. Consider, for instance, a virtual environment simulating ground materials having spatially-varying degrees of compliance, such as gravel areas of different average composition or mud of variable thickness covering a solid floor base: In this case, physically based or physically informed synthesis models such as those mentioned before become unavoidable.

Techniques for the rendering of continuous interactions between objects have reached amazing results in the computer graphics field, thanks to efforts in basic, applied, and technological research leading in particular to recent powerful GPU cores. The audio field, however needing similar research effort and dedicated resources to achieve goals of comparable impact, has come to relatively less resounding results and technologies for its ancillary reputation with respect to vision. Furthermore, computer graphics researchers and developers can rely on the ability of a user to focus

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.)*

on a selected part of the visual field. As opposed to this ability, listeners integrate all simultaneous events of an auditory stream into a single temporal process, whose overall control is known to be almost impossible to achieve in everyday listening environments and ecological soundscapes powered using consumer audio reproduction hardware.

Researchers in interactive synthesis look for models providing accurate sounds and direct access to their distinctive parameters, meanwhile parsimonious in terms of needed computational and memory resources. Successful modeling is usually based on applicable psychoacoustic results, helping sound interaction designers to get rid of redundancies otherwise burdening the resulting synthesis engine. In our specific case, we aim at interactively synthesizing realistic walking sounds by means of a simple model, whose computations and continuous control are at reach of current consumer hardware. Our basic hypothesis, elaborated in the next section, is that perceived material properties in walking sounds map to temporal or spectral auditory cues proportionally to 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. The only experimental variable then will be the floor material.

We categorize floors into *solid* or *aggregate*. The former, such as concrete, marble, wood, are stiff. The latter, such as gravel, dry leaves, sand, allow relative motion of their constituent units and progressively adapt to the sole profile during the interaction. Now,

- solid materials give rise to short, repeatable impacts having a definite spectral color;
- aggregate materials elicit sequences of tiny impacts having distinctive temporal density, that create a sort of "crumpling", less resonant sound.

We hypothesize that the perception of solid materials is mainly determined by *spectral* cues, conversely the perception of aggregate materials is mainly determined by *temporal* cues. 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

A non-interactive setup was used in the experiment. Although forcing the subjects to perform a passive task, this choice was made to keep control on the stimuli and, more in general, focus on the auditory modality. In particular, by avoiding interactive sonic augmentations in the context of a multimodal walking task, subjects could just listen to footstep sounds without incurring in potential bias, determined by self-locomotion over a room floor having completely different visual and tactile attributes.

3.1. Setup

The experiment was set up in the VIPS laboratory at the Department of Computer Science, University of Verona, Italy. Subjects were sitting 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 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 realism and ease of identification of 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 either material, seven recordings were selected and randomly enqueued to create walking sequences lasting 12 s and containing 13 footsteps. In addition to the in-house recordings, high quality samples of a male walking on concrete and on wooden parquet were downloaded from the commercial database `sounddogs.com`. Using these samples, two further walking sequences were created having the same beat and average Sound Pressure Level as of those based on in-house recordings.

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 [5]. By following the input when its magnitude is greater than the envelope, and by in parallel allowing a comparably slow decay of the envelope itself when the same magnitude is smaller, this setting ensures that amplitude peaks are tracked accurately, while leaving spurious peaking components off the envelope signal e_M.

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 manipulated the footstep sequences so to have stationary amplitude along time.

What remained in u_M was a spectral color, that we extracted with a 48th-order inverse LPC filter h_M^{-1} estimated in correspondence of those parts of the signals containing footstep sounds. Using this filter order, if training the model using *one* footstep then we could not detect differences between the original sound and the correspondingly resynthesized footstep. We emphasize that the resulting LPC filter in any case estimated one single transfer function, independently of the number of footsteps taken from the original sequence which informed the model. Since we trained the estimator with the entire sequence, the resynthesized sound had a slightly different color compared to any other footstep belonging to the original sequence.

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

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

This technique draws ideas from a family of physically-informed models of walking sounds [5, 7]. In the meantime it provides a simpler, more controlled resynthesis process avoiding stochastic generation of patterns as in such models. In our case, the silent parts of the four envelopes were tailored to generate synthetic sequences having identical walking tempos. This simple manipulation ensured seamless mutual exchange of the envelopes among sequences, as explained in the following.

Sixteen stimuli were 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 $M_t \neq M_f$:

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

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. This means that the spectral differences in $\tilde{S}_{M_t,M_f}(\omega)$ caused by multiplying $\tilde{u}_{M_f}$ by e_{M_t}, that is,

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

are substantially independent of the material, hence almost identical to those introduced in $\tilde{s}_{M_f}$ by its own envelope e_{M_f}.

Symmetrically, the temporal artifacts which are caused by hybridization between two different materials 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.

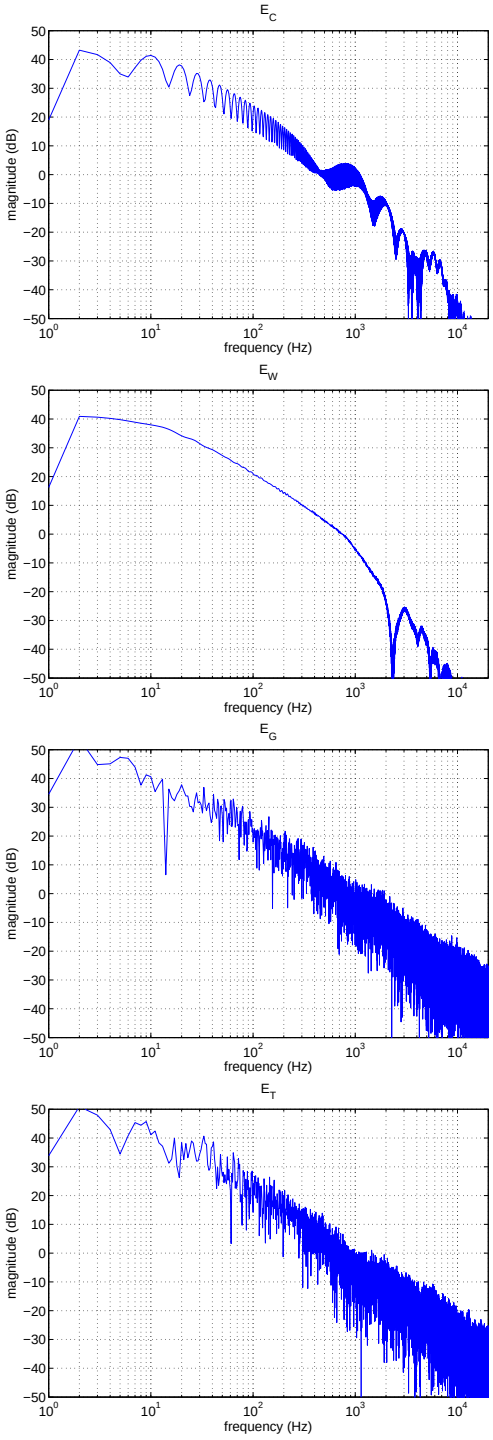

Figure 2: *Magnitude spectra of envelopes E_C, E_W, E_G, and E_T. Respective dc components removed for ease of inspection.*

3.4. Protocol

To become confident with the auditory stimuli and the graphic interface used in the experiment subjects trained for some minutes before starting an individual session, by selecting and playing each one of the original sequences s_C, s_W, s_G, s_T for several times. Figure 3 shows the four graphic buttons exposed by the interface, enabling subjects to play the corresponding original sequences during training.

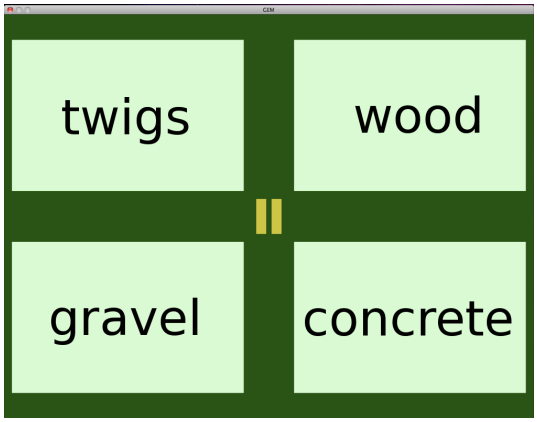

Figure 3: *Translation in English of the graphical interface used in the experiment.*

Each individual session consisted of 192 trials, obtained by randomly playing each one of the sixteen synthetic stimuli for twelve times. At each trial the subject listened to a stimulus, and selected one material by clicking the related button in the interface. When the button was released, the screen froze for two seconds and changed color to inform subjects of the conclusion of the trial. After this short pause, a new trial was performed.

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 45 minutes for each participant to complete the session.

4. RESULTS

For each participant, percentages of selection for the four materials C, W, G, T were analyzed. We considered the fraction of participants who showed significantly correct (random is 25%) material recognitions from the four synthetic stimuli $\tilde{s}_C$, $\tilde{s}_W$, $\tilde{s}_G$, $\tilde{s}_T$, across the twelve repetitions. The critical value (with $\alpha = 0.05$) of the one-tailed binomial test Bin(12,0.25) is 7 trials (i.e., 58.33%): only the participants with an auditory recognition of the original materials higher than 58.33% were included in the analysis. After this check, 16 participants were considered for the recognition of dried twigs, 15 for gravel, 16 for wood, and 10 for concrete.

The results of the analysis are presented in Figure 4. In these plots, a bar exhibiting a low percentage means that the correspondingly substituted 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.

By aggregating the data, we also evaluated the auditory recognition of material categories. Again, this analysis was conducted using data from participants exhibiting an auditory recognition significantly higher than random concerning the two sets of stimuli accounting for the respective categories (24 trials for each category). In this case, the critical value (this time computed by a one-proportion/one-tailed z test to account for the larger number of trials, with $\alpha = 0.05$) is 10 trials, corresponding to 41.67%. All the participants passed the check.

The results of the new analysis are presented in Figure 5. For

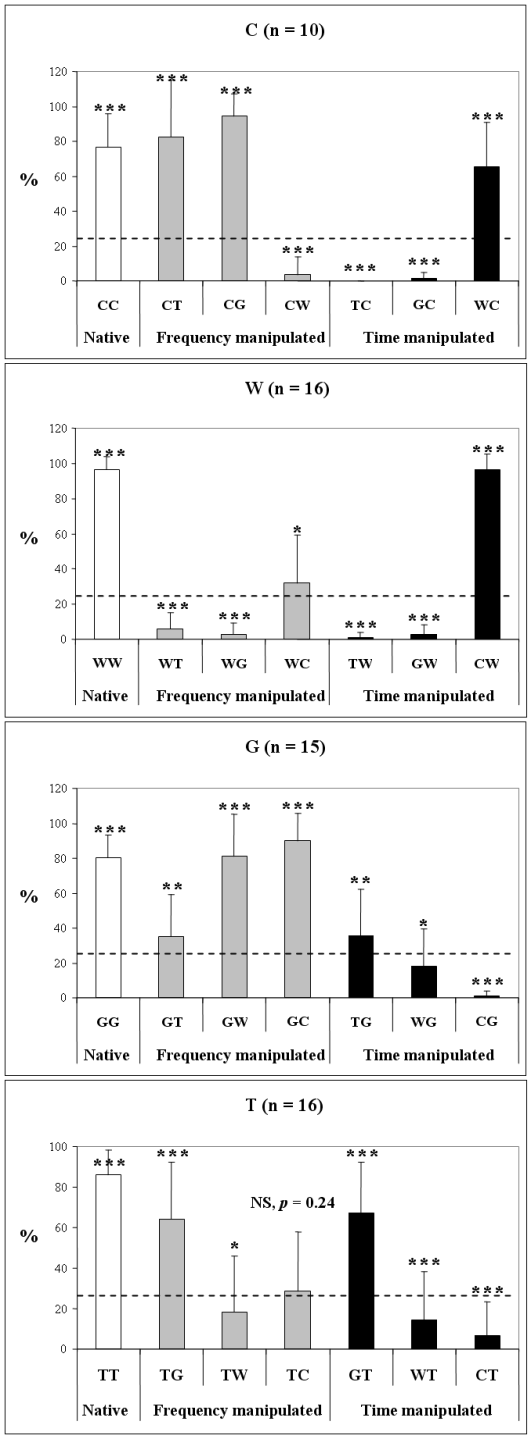

Figure 4: *Mean percentages of selection (lines represent std) for C,W,G,T as a fu nction of the auditory stimulus $\tilde{s}_{M_t,M_f}$. The difference from random selection (line at 25%) was tested using one-proportion (two-tailed) z tests. Note:* $* : p < 0.05$, $** : p < 0.01$, $*** : p < 0.001$,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

the different percentages of selection, the difference relative to ran-

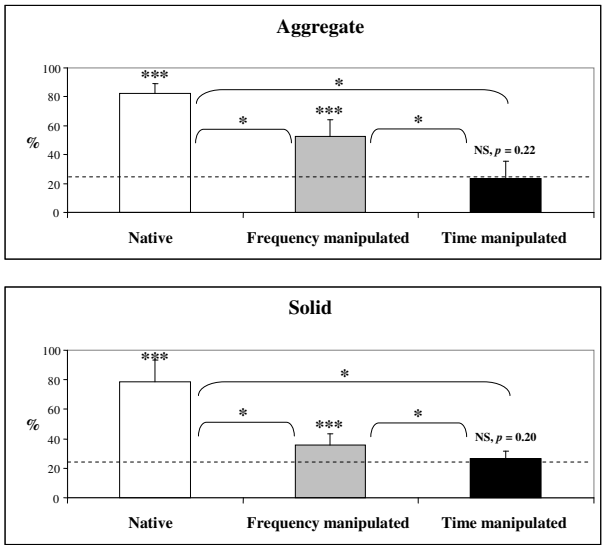

Figure 5: *Mean percentages of selection (lines represent std) for material categories (Aggregate and Solid) as a function of the auditory stimulus $\tilde{s}_{M_t,M_f}$. The difference from random selection (line at 25%) was tested using one-proportion (two-tailed) z tests. The differences between the 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.*

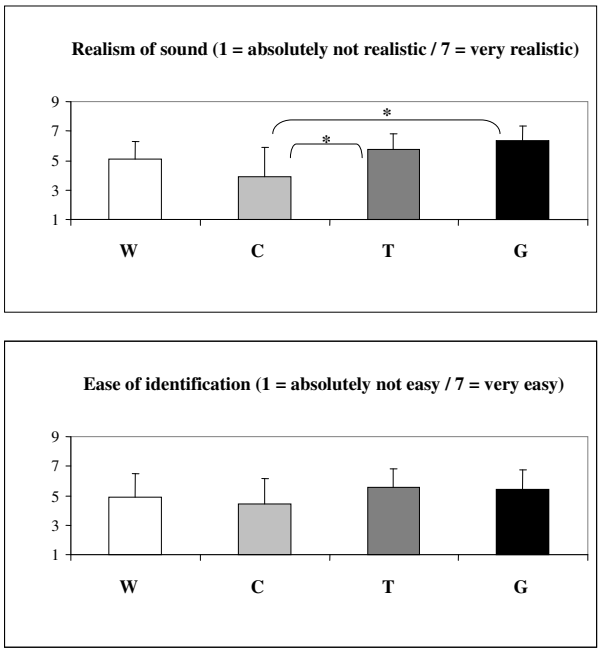

Figure 6: *Mean and standard deviation (lines represent std) of subjective ratings about "realism of sound" and "ease of identification" for the four materials C, W, G, T. Differences between materials were tested with two-proportion z tests (two-tailed and Bonferroni-adjusted alpha level with $p = 0.05/6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

dom (25%) was tested using one-proportion (two-tailed) z tests. Thus, for the aggregate category, the percentages of selection 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 selection 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 selection in native (78.39%) and frequency manipulated (35.59%) conditions were significantly different from random ($z = 24.16$, $p < 0.001$ and $z = 8.30$, $p < 0.001$, respectively). By contrast, the percentage of selection 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, both belonging to the solid category, 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 color 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, spectral color represents an important cue for the recognition of solid materials, conversely sounds of aggregate materials seem to tolerate larger artefacts in their spectra.

Current interfaces operating at ground level, capable of conveying augmented sensations of material properties over otherwise neutral floors, exist in form of active tiles [8] and instrumented shoes [9]. Such interfaces sense the force exerted by a walker over the ground, and instantaneously react by providing auditory and vibrotactile feedback to her or his feet through the action of specific actuators.

The proposed model has been conceived as an alternative to existing real-time physics-based synthesizers onboard those interfaces. Once connected to the physical components realizing their input and output stages, it could easily generate interactive walking sounds by controlling at runtime, through the force sensors and depending on the simulated material, the attack time and shape of the temporal envelope signals, and consequently the amplitude of the LPC-filtered noise forming the auditory and, possibly, also the vibrotactile feedback.

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.*, vol. 123 (5), 2008, 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. 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: AES, Jul. 2002.

[6] A. J. Farnell, "Marching onwards – procedural synthetic footsteps for video games and animation," in *pd Convention*, 2007.

[7] L. Turchet, S. Serafin, and R. Nordahl, "Physically based sound synthesis and control of footsteps sounds," in *Proc. Conf. on Digital Audio Effects (DAFX-10)*, Graz, Austria, Sep. 6-10 2010.

[8] Y. Visell, J. Cooperstock, B. L. Giordano, K. Franinovic, A. Law, S. McAdams, K. Jathal, and F. Fontana, "A vibrotactile device for display of virtual ground materials in walking," in *Proc. of Eurohaptics 2008*, 2008.

[9] S. Papetti, M. Civolani, F. Fontana, A. Berrezag, and V. Hayward, "Audio-tactile display of ground properties using interactive shoes," in *Proc. 5th Int. Haptic and Auditory Interaction Design Workshop*, Sep. 2010, pp. 117–128.