

# Judging the similarity of soundscapes does not require categorization: Evidence from spliced stimuli

Jean-Julien Aucouturier\*

Department of System Studies, The University of Tokyo, Japan

Boris Defreville†

ORELIA, 77300 Fontainebleau, France

This study uses an audio signal transformation, *splicing*, to create an experimental situation where human listeners judge the similarity of audio signals which they cannot easily categorize. Splicing works by segmenting audio signals into 50 millisecond frames, then shuffling and concatenating these frames back in random order. Splicing a signal masks the identification of the categories that it normally elicits: for instance, human participants cannot easily identify the sound of cars in a spliced recording of a city street. This study compares human performance on both normal and spliced recordings of environmental ambiences (or soundscapes) and music. Splicing is found to degrade human similarity performance significantly less for soundscapes than for music: when two spliced soundscapes are judged similar to one another, the original recordings also tend to sound similar. This establishes that humans are capable of reconstructing consistent similarity relations between soundscapes without relying much on the identification of the natural categories associated with such signals, such as their constituent sound sources. This finding contradicts previous literature and points to new ways to conceptualize the different ways in which human perceive soundscapes and music.

PACS numbers: 43.66.Ba, 43.50.Rq, 43.75.Cd, 43.60.Cg

Keywords: Splicing, Soundscapes, Music, Cognition

## I. INTRODUCTION

Categorizing and comparing percepts are central abilities for human cognition. They do not operate in isolation, but rather depend greatly on one another. When we lack specific knowledge, we often rely on domain-general similarity to guide categorization<sup>1</sup>. Reciprocally, we often judge the similarity of objects by comparing their features, i.e. based on their prior categorization<sup>2</sup>. However, the exact nature of the relation between similarity and categorization still eludes modern cognitive science<sup>3</sup>. There have been empirical demonstrations of categorization processes conducted in apparent independence from similarity, but based on semantic representations looking like rules or theories<sup>4</sup>. Further, features used for similarity depend on context, culture and experience, and different strategies often operate at the same time to reinforce each other and eliminate context-driven noise.

This complex interdependence also holds in the specific case of auditory perception. The vast psychology literature studying similarity judgments for musical signals has made constant efforts to control for unwanted high-level recognition effects, by using short or synthesized stimuli<sup>5,6</sup>. Research on the perception of environmental audio scenes (or *soundscapes*) also suggests that similarity and categorization are deeply entangled. The perceived similarity of soundscapes is believed to be based

on the prior identification of their constituent sources<sup>7,8</sup>: “if there are cars in both signals, then these are similar”. Similarly, psycholinguistic studies<sup>9</sup> have concluded that the subjective “unpleasantness” of a urban soundscape increases when more mechanical sound sources (e.g. vehicles) are identified than natural sources (e.g. voices or birds).

In the case of soundscapes however, recent computational studies seem to contradict the apparent dependency of similarity on categorization. A common paradigm to simulate human judgments of audio similarity algorithmically is to model audio signals as a global statistical distribution of local features computed on very short (c. 50ms) frames. This technique, called bag-of-frames (BOF), uses only low-level and short-term spectral information, and does not rely on any prior categorization of e.g. sound sources. In a recent study<sup>10</sup>, we established that BOF, simple as it is, was computationally sufficient to simulate human judgments of audio similarity with near-perfect precision for soundscapes. However, we found significantly lower performance when applied to polyphonic music signals. Therefore, from the sole point of view of information processing, categorization does not seem to be necessary to compare soundscape signals.

The present paper aims to clarify this situation. While our previous study<sup>10</sup> was a machine experiment, this one is a human psychological experiment. We create an experimental situation where human listeners produce judgments of similarity for audio signals which they cannot easily categorize, by manipulating the signals with an algorithmic transformation called *splicing*. Splicing alters an audio signal by segmenting it into 50 millisecond

---

\*Electronic address: [aucouturier@gmail.com](mailto:aucouturier@gmail.com)

†Electronic address: [boris.defreville@orelia.fr](mailto:boris.defreville@orelia.fr)

frames, then shuffling and concatenating these frames back in random order. Spliced signals by construction have the same frames than the original signals, hence the same local features, as well as the same statistics of features; only the time ordering of the frames is modified.

Splicing is a known technique in language science: it is typically used to process spoken sentences for listening experiments so that their lexical content is lost but their vocal content (prosody, timbre, etc.) preserved to some extent<sup>11</sup>. Similar effects are known for music signals: it is difficult to identify instruments or even produce any qualitative description of shuffled or reversed music<sup>12</sup>. This is hardly surprising: psychophysical experiments<sup>5</sup> have long established the importance of dynamics, notably the attack time and fluctuations of the spectral envelope, in the perception of individual instrument notes.

In this study, we confirm that splicing acts as a categorization mask for both music and soundscape, and use this property to study the influence of categorization on the processing of similarity for music and soundscapes.

## II. METHODS

### A. Data

Each subject was presented (in tasks described below) a set of 12 audio stimuli in either of 4 conditions: normal music (NM), spliced music (SM), normal soundscapes (NS), spliced soundscape (SS). Each stimuli is a 20-second stereo sound file sampled at 44,1kHz, normalized at zero DC and maximum amplitude -3dB. The 12 stimuli in the NM condition are excerpts of western popular music of various genres from the period 1950-2000. The 12 stimuli in the NS condition are excerpts of urban soundscapes used in our previous study<sup>10</sup>, recorded in Paris using an omni-directional microphone. Stimuli in the 2 spliced conditions (SM,SS) were obtained from the stimuli in the associated normal condition (resp. NM,NS) by the following algorithmic transform:

1. a “normal” signal is truncated into successive, non-overlapping 2048-point time frames (46,4ms at 44,1kHz)
2. frames are shuffled in random order (using a pseudo-random generator initialized with a common seed for all files)
3. frames are concatenated back to an audio signal of same length as original

### B. Participants

55 subjects participated, all relatively young ( $M = 20.70, SD = 3.56$ ) Japanese undergraduate students. There were 45 male participants, and 10 female participants, with no significant age difference between sexes

( $F(1, 55) = 1.15, p = 0.288$ ). All participants distributed in roughly equal proportion over all 4 conditions (NM:15,SM:15,NS:13,SS:12), with no gender bias. We kept data from all participants for the analysis presented below.

### C. Tasks

Each subject was presented stimuli from only 1 of the 4 tested condition (randomized). The test consisted of a categorization task (used to validate the effect of splicing) and a similarity task. To prevent priming effects, every subject performs the similarity task before the categorization task.

#### 1. Similarity Task

In the similarity task, participants were asked to judge similarity relations in a set of triads presented one after the other. We presented 44 triads per condition, arranged in a  $\lambda = 2$  balanced incomplete block design (all possible pairs of stimuli appear twice). All triads in a given spliced condition have a corresponding triad in the corresponding normal condition composed of the 3 same stimuli in their original version. For each presented triad, participants were asked to select the pair of stimuli they deemed were most similar to one another.

participants were given no other information than the sole audio presentation of the stimuli (no file name, no editorial information on the recording, etc.). Prior to the test, participants received the following instructions:

When judging the “similarity” of such sounds, please try to listen to the immediate, global impression of the sound, and not to finer, analytical details<sup>13</sup>. For instance, you may want to say that a Bob Dylan song “sounds like” another piece (e.g. a Bossa Nova piece by Joao Gilberto), because they both have a soft acoustic guitar, and a gentle male voice. In a noisy background, e.g. a crowded cafe, one could be mistaken for the other. Judging that the same Bob Dylan guitar song is similar to, say, a Joni Mitchell piano song (because both artists have the same influences, and performed at similar periods), or that the Bossa Nova piece is similar to a recent Brazilian electronic piece (because they’re both from Brasil) would be irrelevant for this test.

Additionally, when tested in one of the spliced conditions (SM,SS), participants were instructed to

[...] please try to listen to these sound extracts as if they were natural ones, i.e. focus on your immediate, holistic sensation of how

much these sounds "sound like" one another (and try to refrain from analytical thinking).

Stimuli in each triad could be listened to as many times as needed; answers could be given before listening to a stimuli in full; in case they could recognize the original recording from which a given extract was taken (which is mostly relevant for the music conditions), participants were instructed to judge similarities based on the sole extract, and not on the unrepresented original material.

## 2. Categorization Task

In the second task, each of the 12 stimuli of the tested condition are presented again one after the other. For each stimuli, users were instructed to check through a list of descriptions that they think could apply (forced choice). As before, participants were given no other information than the sole audio presentation of the stimuli. The presented descriptions were organized in 3 broad categories (sources, moods, types), which content varied depending on the tested condition (full list given in Appendix 1).

Each category was designed to contain descriptions that apply naturally for all items (e.g. "car" sounds for a "boulevard" soundscape) as well as distractors which do not apply<sup>14</sup> for any of the presented data (e.g. "violin" source for music, "boat" sounds for soundscapes). In each category, a wildcard description ("I don't know") was also included. Apart from the wildcard choice, none of the descriptions were enforced to be mutually exclusive, i.e. participants were left free to check as many as they thought could apply. Prior to the test, participants received the following instructions:

This time, the task may require you to think analytically about the signals: which instruments/sound sources you perceive, how you judge the sound emotionally, how you would describe the genre of the music, or to which type of environment the recording corresponds.

Additionally, when tested in one of the spliced conditions (SM,SS), participants were instructed to

[...] please try to describe what you think the signal was *before* scrambling. For instance, all scrambled signals tend to sound rhythmically complex and maybe evoke tensed emotions associated with randomness. Please try to refrain from judging such properties, but rather focus on what you can perceive of the original signal. Imagine you're listening to the sound over a very strangely distorted telephone: we want you to describe the sound played at the other end, not to describe the telephone.

As before, each successive stimulus could be listened to as many times as needed; answers could be given before listening to a stimulus in full; participants were instructed to categorize the content of the sole extracts, rather than any unrepresented context they might recognize.

## D. Apparatus

Data was gathered in 4 sessions held on 4 days of December 2007, with randomized conditions. All participants were tested in the same room, in a calm environment, with the same apparatus: audio was presented with closed-type headphones (always the same model), and the test application (a set of php scripts running on a local server) ran on G4 Apple computers. All instructions and tests were conducted in Japanese. Participants were instructed to set the audio volume to a comfortable level before the test. 3 experimenters were present at any time to give instructions and answer queries. Participants were rewarded 1,000 JPY (c. 10 USD) for their participation which took between 30 and 60 minutes.

## III. RESULTS

### A. Influence of splicing on categorization

In this section, we measure how much the descriptions (sources, moods, types) that apply consensually to the normal stimuli can still be judged consistently in the corresponding spliced stimuli.

For a given stimulus  $s$  in a given condition  $c$ , we define the "agreement on best value" for a given description  $d$   $\beta_c(s; d)$  as the proportion of participants who agree on the most-agreed value (i.e. true or false) of description  $d$  for  $s$ . For example, when presented an extract of a *Beethoven* piano sonata, most participants agree that the value for the description *piano* (a member of the *source* is "true". For stimuli  $s$  in a *spliced* condition  $c$  (i.e. one of SM and SS), we then compute the "agreement on best normal value"  $\beta_c^*(s; d)$  as the proportion of participants (among all participants tested in condition  $c$ ) who agree on the most-agreed value (i.e. true or false) of description  $d$  judged for the corresponding normal stimuli. We then average over all stimuli  $s$  in condition  $c$ , then all descriptions  $d$  in a given category (i.e. one of "source", "mood" and "type"). With the same example as above,  $\beta_{SM}^*(s; d)$  measures the proportion of participants judging description *piano* to be "true" for the spliced variant of the sonata.

We observe that:

- There is a severe and significant degradation of the agreement measured in spliced conditions over the most-agreed values judged in the corresponding normal conditions, when considering only TRUE values in normal condition, i.e. increased number of false negatives (see Table I).

- No significant degradation of agreement due to splicing when considering only FALSE values in normal condition, i.e. no increase of false positives.
- Splicing reduces the number of hits per (stimulus,description) pair (participants check less descriptions for spliced signals than for the corresponding signals in normal condition), for all category groups and for both music and soundscape, but only significantly so for soundscape sources (number of checks:  $M = 2.75 < M = 1.89$ ,  $F(1,19)=7.93$ ,  $p=0.011$ )
- The degradation of  $\beta_c^*$  due to splicing on the categorization tasks is of similar degree for music and soundscapes, for the mood category ( $\Delta = -0.3$ , but  $F(1,47)=0.02$ ,  $p=0.891$ ) and the type category ( $\Delta = -0.28$ , but  $F(1,14)=0.01$ ,  $p=0.944$ ). The degradation of the agreement on source values is more severe for soundscapes than music: the agreement difference from NM to SM  $\beta_{SM}^*(s; d \in source) - \beta_{NM}(s; d \in source)$  ( $M = -0.20$ ) is smaller than from NS to SS  $\beta_{SS}^*(s; d \in source) - \beta_{NS}(s; d \in source)$  ( $M = -0.37$ ), with statistical significance ( $F(1,68)=7.48$ ,  $p=0.008$ )

These results show that splicing significantly hinders the recognition of sources, moods and types consensually associated (in the tested control groups) with stimuli in their normal conditions. This masking effect is not significantly different for music and soundscapes for both mood and type descriptions, but more important for soundscape sources than for music sources.

## B. Influence of splicing on similarity performance

In this section, we measure how much of the similarity relations established by participants listening to normal stimuli (“ground truth”) can still be identified by the participants listening to their spliced variants.

For a given triad  $t$  in a condition  $c$ , we define the “agreement on best pair”  $\alpha_c(t)$  as the proportion of participants (among all participants tested in the condition) who agree on the most-agreed pair of “most” similar stimuli. For example, when presented a triad composed of 2 *Beethoven* piano sonata and a *Coltrane* jazz piece, most participants judge that the pair of most similar stimuli is the one composed of the 2 sonata. For triads  $t$  in the spliced conditions  $c$ , we then compute the “agreement on best normal pair”  $\alpha_c^*(t)$  as the proportion of participants (among all participants tested on spliced condition  $c$ ) who agree over the most consensual pair judged by the participants tested on the same triad, but in the normal condition. We then average over all triads in the condition. With the same example as above,  $\alpha_{SM}^*(t)$  is the proportion of participants who, when presented the triad composed of the spliced variants of the 2 sonata and the jazz piece, judge that the pair of most similar stimuli is the one composed of the 2 spliced sonata.

We observe that:

- There is a significant degradation of agreement in the spliced conditions over the pairs most-agreed on in the normal conditions, i.e. if 2 stimuli are consensually judged similar to one another in their normal condition, their spliced variants are not necessarily judged similar. This is true<sup>15</sup> for music ( $\alpha_{NM} = 0.86 > \alpha_{SM}^* = 0.53$ ,  $F(1,48)=6.78$ ,  $p=0.012$ ) as well as soundscapes ( $\alpha_{NS} = 0.91 > [\alpha_{SS}^*] = 0.77$ ,  $F(1,48)=6.28$ ,  $p=0.016$ )
- However, the degradation due to splicing is 2.5 times more severe for music than for soundscapes: the agreement difference  $\alpha_{NM}(t) - \alpha_{SM}^*(t)$  ( $M = 0.33$ ,  $SD = 0.27$ ) is larger than  $\alpha_{NS}(t) - \alpha_{SS}^*(t)$  ( $M = 0.13$ ,  $SD = 0.25$ ), with statistical significance ( $F(1,44)=6.38$ ,  $p=0.015$ )

Spliced soundscapes have only slightly different similarity relations to one another than in the normal condition, whereas the relations consensually judged in normal music are lost when it is spliced. This opposes all effects found on categorization, which is equally hindered for music and soundscapes (and even more for soundscapes in the case of “source” categories).

## IV. DISCUSSION

We found that splicing significantly impedes the recognition of source, mood and type descriptions that can be consensually made (by our tested participants) for stimuli in their normal condition. We also found splicing has a significant, but small influence on similarity performance for soundscape signals; for music however, the similarity relations that are consensually established for normal signals are completely lost by splicing.

The difference observed between music and soundscape signals when spliced may result from several undesired effects. First, similarity judgments could be more difficult (i.e. less consensual) for music than for soundscapes even in the normal condition. However, we found that similarity judgments for normal music signals are as consistent across participants as judgments for normal soundscape signals. ( $\alpha_{NM} = 0.70 < \alpha_{NS} = 0.77$ , but  $F(1,86)=3.34$ ,  $p=0.071$ ), so this effect is ruled out.

Second, judging the similarity of spliced music maybe be a more disorienting task than for spliced soundscapes. But again, we found that similarity judgments for spliced signals are as consistent as judgments for normal signals ( $\alpha_{NM} = 0.70 > \alpha_{SM} = 0.66$  ( $F(1,86)=1.27$ ,  $p=0.262$ );  $\alpha_{NS} = 0.77 > \alpha_{SS} = 0.76$  ( $F(1,86)=0.13$ ,  $p=0.719$ ). Poor similarity precision on spliced music is *consistently* observed for all participants.

Third, our results may be explained by the participants’ lack of training for the task of listening to spliced signals. While a positive effect of training cannot be ruled out, it is interesting to note that judg-

ments made on spliced signals reach the same degree of consensus than judgments made on ecologically valid, normal signals. For instance, the categories judged for spliced stimuli are as consistent as the categories for normal signals, e.g. for music moods ( $\beta_{NM}(mood) = 0.90 < \beta_{SM}(mood) = 0.90$ ,  $F(1,26)=.03$ ,  $p=0.873$ ), types ( $\beta_{NM}(type) = 0.85 < \beta_{SM}(type) = 0.85$ ,  $F(1,22)=.00$ ,  $p=0.998$ ) or sources ( $\beta_{NM}(source) = 0.86 < \beta_{SM}(source) = 0.83$ ,  $F(1,29)=.35$ ,  $p=0.561$ ).

Finally, the difference observed between music and soundscapes may result from uncontrolled recognition effects of familiar music. If participants recognize the music stimuli used in the experiment, they may judge music similarity in the normal condition based on extra-acoustical information such as the artists' country. Since this recognition is likely lost when spliced, this information cannot be recruited to compare spliced music, and performance is degraded. For soundscapes however, this recognition effect is negligible, which could explain the lesser difference from normal to spliced. This explanation can probably be ruled out, too. First, we found in debriefing the experiments that the participants did not recognize the stimuli in the normal music condition (pieces that are relatively well-known in the western world are not necessarily well-known in Japan). Second, it was recently reported that listeners can make a distinction between familiar and unfamiliar music even in spliced stimuli with frame size between 50 and 100ms.<sup>16</sup>

### A. Splicing and syntax

Spliced signals sound distinctively unnatural. Physically, splicing preserves long-term statistical distributions of frames as well as local spectral content (since each individual frame is left untouched), but loses short-term dynamical relationships between frames. In particular, the frame rate adopted here is typically faster than the note rate in music, so even intra-note consistency (transient, sustain, decay) is not preserved. Brain imaging studies show that spliced musical signals (for frame size around 500ms) engender decreased activation in the left inferior frontal cortex<sup>17,18</sup>, a region closely associated with the processing of linguistic structure, or *syntax*, in spoken and signed language.

We found here that humans are capable of comparing soundscapes in a way which resists splicing: two soundscapes deemed similar to one another in normal condition are likely to be also found similar when listened in spliced condition. This suggests that judging the similarity of soundscapes does not require to recruit syntactic processing of the type involved by speech and music (Figure 1-left). Whether syntactic processing is activated by the perception of non-spliced soundscapes but contributes little, or whether it is simply not recruited even in normal situations, is left open for further investigation.

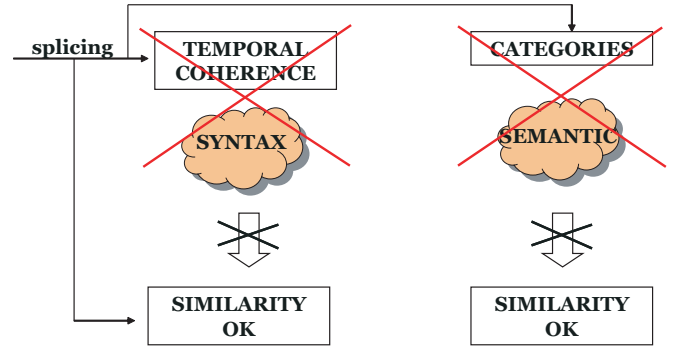


FIG. 1. (color online) A schematic representation of the links found here between similarity, syntax and semantics for soundscape perception. Splicing affects the temporal coherence of the audio signals, which was found to hinder activation in brain centers responsible for the processing of linguistic syntax. Since splicing preserves soundscape similarity to a large extent, we find here that soundscape similarity does not logically require syntactic processing of the type involved by speech and music (left). Moreover, splicing is found to hinder categorization. Therefore, soundscape similarity does not require the identification of the categories usually elicited by normal stimuli, most notably sound sources. (right). See main text for details.

### B. Splicing and semantics

Our results confirm that splicing is a categorization mask for non-speech audio signals. Degraded agreements for “true” categorical values, but not for “false” values, show that splicing masks the recognition of normally perceived categories (increased number of false negatives), but does not introduce any confabulated categories (no significant increase of false positives): one may not hear guitar in a spliced guitar signal, but one is unlikely to hear piano instead. This effect is the same for music and soundscapes.

These results are consistent with the use of splicing for speech signals<sup>11</sup>, and can be easily explained by the well-documented importance of the temporal envelopes for, say, musical instrument recognition<sup>5</sup>. However, they appear to contradict studies focusing on the perception of very short sounds. In a 1999 study (published only recently), Gjerdingen and Perrott<sup>19</sup> observed that, given as little as 250 ms of audio, human listeners were able to recognize the genre of a musical extract correctly more than 70% of the times. Similar effects were reproduced for further categories (including moods) for extracts as small as 50ms<sup>20</sup> - which is the size of the frames we use here. That this ability should not hold for *sequences* of frames is intriguing and could point to an important role of short-term memory and attentional processes.

From our data, it is difficult to clarify the causal relation between loss of temporal coherence (by construction) and distorted ability to categorize. Friend and Farrar<sup>21</sup> studied the effect of splicing speech on judgments of the speaker’s affective state, and found that it

increased ratings of anger and excitement. They suggested this depends on the speech-mimetic features of the resulting time series. Spliced signals contain regular artificial transients (due to the arbitrary concatenation of non-adjacent frames) and give an impression of high tempo and limited pausing regardless of the original signal. These features are often correlates of anger in normal speech. For music and soundscapes, we observed here that splicing typically did not introduce confabulated categories, and we could not find any systematic effect of splicing on the categorization of “mood” descriptions. However, similar effects cannot be ruled out and should be further investigated.

### C. Soundscape similarity doesn’t require categorization

The better similarity results obtained in spliced condition for soundscapes than for music could be explained by a possible recognition effect which would be retained in the case of soundscapes. If more analytical details (i.e. descriptions) were preserved in spliced soundscapes than in spliced music, then their comparably better similarity could be based on a comparison of high-level features in the way proposed by Tversky<sup>2</sup>. However, we found that splicing degrades categorization to a similar degree for both music and soundscape, and even to a greater degree for soundscapes when considering sound sources, so this effect can be ruled out. The fact that splicing preserves more of the music sources (instruments) than the soundscape sources would in effect predict better similarity performance for music than for soundscapes, if based on analytical features. Moreover, we find here that relatively good soundscape similarity doesn’t require source, mood or type categorization (we have the former without the latter in the spliced condition) (Figure 1-right).

First, this suggests that a significant part of the process of comparing two soundscapes doesn’t require high-level semantic features, but rather relatively raw representations, arguably generic. Second, this result indicates in particular that humans are able to give accurate and consensual judgments of the similarity of soundscapes with no prior recognition of their constituent sound sources. This is a surprising result, in apparent contrast with previous studies which usually conclude that perceptive judgments of categorization and similarity on soundscapes are based on the identification, and then connotation, of sources. Psycholinguistic studies<sup>9</sup> only report on correlations. There are well-studied effects of “a posteriori rationalization” in such methodologies<sup>22</sup>, in which participants produce a lexical explanation (“*because* there are cars”) for a cognitive process (here, “these two soundscapes are similar”) which may be independent from these causes.

We should be careful in concluding too strongly on “similarity without categorization”. It is probably impossible to prevent categorization in normally cognitive participants, even for highly distorted stimuli such as the

ones used here. What we are losing here by splicing is referential categorization in the conceptual system associated with normal stimuli: participants find it impossible to map a spliced stimulus to categories that apply for normal stimuli, such as “piano” or “dog bark”. However, when presented a spliced stimulus, participants likely build categories in a ad-hoc semantic system adapted to spliced sounds (e.g. “sounds that go *flpflpflp*”). What we find here is that these ad-hoc categories bear no relation with the “natural” categories, as this spontaneous categorization of spliced signals doesn’t help the participants retrieve the categories consensually judged in the normal conditions. It is possible yet, if properly trained with pairs of normal and spliced stimuli, that participants can learn a mapping between both semantic systems, e.g. learn that “sounds that go *flpflpflp*” are usually spliced variants of songs containing *guitar*.

### D. Does music similarity require categorization ?

We found here that severely degraded music similarity (-33% agreement) is correlated with severely degraded categorization (c. -30% agreement). From this data, we cannot conclude on the causality between these 2 processes. First, maybe similarity requires categorization, i.e. two music stimuli that are said to “sound the same” can only be said so because they are first categorized with a variety of descriptors (e.g. musical instruments, moods, etc.) and then found to share a significant number of such descriptors. If this is so, then splicing acts as a content-mask on categorization, and impedes similarity by hurting its prerequisite process. Conversely, maybe categorization requires similarity, i.e. music classification operates as an exemplar matching process<sup>23</sup> based on a partly generic assessment of similarity (e.g. if it sounds like a rock song, then it is rock). The content-masking property of splicing could then be the consequence, rather than the cause, of this degraded similarity. Finally, maybe music categorization and similarity operate in conditional independence given a common cause, such as syntactic processing, which is degraded by splicing. This should be submitted to further investigation, e.g. with the help of brain imaging techniques which are increasingly used to characterize semantic processing both for music<sup>24</sup> and natural sounds<sup>25</sup>.

### Acknowledgments

The authors would like to thank Brian Gygi, Geraint Wiggins and anonymous reviewers for their useful comments. This research was partially funded by a JSPS Postdoctoral Fellowship for Foreign Researcher. Yusuke Kanayama helped with the logistics of the experimentation and Masumi Nasukawa with the Japanese translation of our instructions.

## APPENDIX: DESCRIPTIONS USED IN CATEGORIZATION TASK

Descriptions for normal and spliced music:

- Sources: Acoustic Guitar, Bass, Brass (trumpet...), Drums, Electric Guitar (saturated), Electric Guitar (clear), Harmonica, Piano, Strings (violin...), Synthesizer, Vocals (male), Vocals (female), Wind (flute...).
- Moods: Aggressive, Calm, Carefree, Cerebral, Cold, Energetic, Fun, Happy, Intimate, Party / Celebratory, Passionate, Plaintive, Sad, Tense/Anxious.
- Types: Blues, Club / Dance, Classic Rock & Roll, Contemporary Rock, Country, Electronic, Experimental, Folk, Hard Rock, Heavy Metal, Jazz, Pop, Punk, Psychedelic.

Descriptions for normal and spliced soundscapes:

- Sources: Car, Truck, Bus, Bird, Voice (male), Voice (female), Voice (children), Boat, Motorbike, Footsteps, Plane, Train, Industrial machinery, Building site equipment, Dog bark, Water noise (rain, sea), Collision noise, Car horns / siren.
- Moods: Calm, Disturbing, Familiar, Noisy, Pleasant, Restful, Stimulating, Stressful, Unbearable, Unpleasant.
- Types: Airport, Avenue, Calm street, Construction site, Factory, Highway / Motorway, Inter-city road, Outway for emergency vehicles, Park, Railway, School yard, Shopping/Pedestrian street, Street market, Technical equipment.

- [1] R. M. Nosofsky, "Attention, similarity and the identification-categorization relationship", *Journal of Experimental Psychology: General* **115**, 39–57 (1986).
- [2] A. Tversky, "Features of similarity", *Psychological Review* **84**, 327–352 (1977).
- [3] R. L. Goldstone and J. Son, *Similarity*, 13–36 (Cambridge: Cambridge University Press) (2005).
- [4] G. L. Murphy and D. L. Medin, "The role of theories in conceptual coherence", *Psychological Review* **92**, 289–316 (1985).
- [5] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes", *Psychological Research* **58**, 177–192 (1995).
- [6] T. Eerola, T. Jarvinen, J. Louhivuori, and P. Toivainen, "Statistical features and perceived similarity of folk melodies", *Music Perception* **18**(3), 275–96 (2001).
- [7] J. Ballas, "Common factors in the identification of an assortment of brief everyday sounds", *Journal of Experimental Psychology, Human Perception and Performance* **19**, 250–267 (1993).
- [8] V. Peltonen, A. Eronen, M. Parviainen, and A. Klapuri, "Recognition of everyday auditory scenes: Potentials, latencies and cues", in *Proceedings of the 110th Convention of the Audio Engineering Society, Amsterdam (The Netherlands)* (2001).
- [9] D. Dubois, C. Guastavino, R. Maffiolo, and M. Raimbault, "A cognitive approach to soundscape research", *Journal of the Acoustical Society of America* (Invited paper at the 147th (Acoustical Society of America) meeting) **115**, 2592 (2004).
- [10] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music", *Journal of the Acoustical Society of America* (accepted) (2007).
- [11] K. R. Schere, "Randomized splicing: A note on a simple technique for masking speech content", *Journal of Experimental Research in Personality* **5**(2), 155–159 (1971).
- [12] M. Castellengo and D. Dubois, "Timbre ou timbres? propriétés du signal, de l'instrument ou construction cognitive?", in *Proceedings of the 2005 Conference on Interdisciplinary Musicology (CIM05), Montreal, Canada* (2005).
- [13] With this instruction, we aim to direct participants toward holistic rather than selective listening - these 2 strategies have been found to induce significantly different brain signatures, see Janata, P., Tillmann, B. and Bharucha, J. Listening to polyphonic music recruits domain-general attention and working memory circuits. *Cognitive, Affective and Behavioral Neuroscience*, 2(2), 121-140 (2002).
- [14] Or at least were not expected to apply consensually.
- [15] The values reported here discard outlier triads with particularly low consensus: 7 triads for condition NM, 4 triads for condition NS.
- [16] B. Tillmann and E. Bigand, "Global context effects in normal and scrambled musical sequences", *Journal of Experimental Psychology: Human Perception and Performance* (2008).
- [17] D. J. Levitin and V. Menon, "Musical structure is processed in 'language' areas of the brain: A possible role for brodmann area 47 in temporal coherence", *NeuroImage* **20**, 2142–2152 (2003).
- [18] T. Matsui, K. Kazai, M. Tsuzaki, and H. Katayose, "Investigation of the musician's brain activation during different music listening modes: A near-infrared spectroscopy study", in *Proceedings of the 2008 International Conference on Music Perception and Cognition, Sapporo, Japan* (2008).
- [19] R. Gjerdingen and D. Perrott, "Scanning the dial: The rapid recognition of music genres", *Journal of New Music Research* **32**(7) (2008).
- [20] R. Ashley, "Affective and perceptual responses to very brief musical stimuli", in *Proceedings of the 2008 International Conference on Music Perception and Cognition, Sapporo, Japan* (2008).
- [21] M. Friend and M. J. Farrar, "A comparison of content-masking procedures for obtaining judgments of discrete affective states", *Journal of the Acoustical Society of America* **96**, 1283–1290 (1994).
- [22] P. Johansson, L. Hall, S. Sikström, and A. Olsson, "Failure to detect mismatches between intention and outcome in a simple decision task", *Science* **310**, 116–119 (2005).
- [23] R. M. Nosofsky, "Exemplar-based accounts of relations between classification, recognition and typicality", *Jour-*

- nal of Experimental Psychology: Learning, Memory and Cognition **14**, 700–708 (1988).
- [24] S. Koelsch, E. Kasper, D. Sammler, K. Schulze, T. Gunter, and A. Friederici, “Music, language and meaning: Brain signatures of semantic processing”, *Nature Neuroscience* **7(3)** (2004).
- [25] A. Cummings, R. Ceponiene, A. Koyama, A. P. Saygin, J. Townsend, and F. Dick, “Auditory semantic networks for words and natural sounds”, *Brain Research* **1115(1)**, 92–107 (2006).

TABLE I. Influence of splicing manipulation on inter-subject agreement on most-agreed value for categorization task

		Normal condition	Spliced condition	Significance
music	source	$\beta_{NM} = 0.79$	$\beta_{SM}^* = 0.59$	$F(1, 80) = 18.23, p < 0.001$
	mood	$\beta_{NM} = 0.67$	$\beta_{SM}^* = 0.37$	$F(1, 62) = 73.05, p < 0.001$
	type	$\beta_{NM} = 0.64$	$\beta_{SM}^* = 0.36$	$F(1, 8) = 10.56, p = 0.012$
soundscapes	source	$\beta_{NS} = 0.89$	$\beta_{SS}^* = 0.52$	$F(1, 56) = 30.19, p < 0.001$
	mood	$\beta_{NS} = 0.68$	$\beta_{SS}^* = 0.38$	$F(1, 32) = 28.70, p < 0.001$
	type	$\beta_{NS} = 0.75$	$\beta_{SS}^* = 0.46$	$F(1, 20) = 6.75, p = 0.017$