

ICMPC10 Spoken presentation; empirical research

## **DIFFERENCES IN THE COGNITIVE PROCESSING OF MUSIC AND SOUNDSCAPES REVEALED BY PERFORMANCE ON SPLICED STIMULI**

*Jean-Julien Aucouturier*

The University of Tokyo, Japan

### **BACKGROUND**

Recent years have seen considerable effort in the pattern recognition community to simulate human auditory perception computationally. Most algorithms to this aim rely on a common paradigm, so-called bag-of-frames (BOF), which models signals as a global statistical distribution of local features computed on ~50ms frames. BOF algorithms typically yield near perfect performance on tasks involving sonic environments (soundscapes), but only limited precision for polyphonic music.

### **AIMS**

It is often argued that such discrepancy results from physical differences between both types of signals, music requiring more advanced features and/or statistical modelling. This study aims at testing a different hypothesis, namely that the results of BOF algorithms on music are perfect, but for a wrong task: the information they are provided with may be insufficient even for human cognition.

### **METHOD**

BOF algorithms, which typically take no account of time, listen in effect to spliced audio signals, i.e. signals which frames have been shuffled randomly in time and then concatenated back. We compare human performance on normal and spliced signals, for both music and soundscapes, in 2 tasks: a triadic similarity test and a forced-choice categorization. Additionally, we compare human performance with that of a typical BOF algorithm, which is by construction identical in both normal and spliced condition.

### **RESULTS**

First, we find that splicing significantly degrades human similarity performance, both for music and soundscapes, but significantly more so for music than soundscapes. Second, we observe that human similarity performance on spliced music signals is no better, and sometimes even worse, than the performance of BOF machines – while it is better than machines for spliced soundscapes. Finally, we find that splicing severely degrades categorization both for music and soundscapes, but, contrary to similarity performance, with no significant difference in degree between both types of signals.

### **CONCLUSIONS**

Our results suggest fundamental differences in the cognitive processing of music and soundscapes. On the one hand, humans perform very poorly on the kind of musical data we typically expect BOF algorithms to succeed on - in fact they're even worse than machines. This suggests that the next

## ICMPC10 Spoken presentation; empirical research

frontier in reaching human-level music processing is beyond the assumptions of BOF models, which are near-perfect solutions to a wrong (although difficult) problem.

On the other hand, the BOF assumption for soundscapes is not only computationally effective, but also cognitively sufficient: it appears that humans are capable of comparing soundscapes in a timeless, amorphous way which resists splicing.

Moreover, good similarity performance on soundscapes doesn't appear to require identification of e.g. constituent sound sources, which indicates that soundscapes can be compared in an acoustic-only manner, without much semantic analysis. Conversely, semantic analysis may be precisely what splicing impedes, and BOF algorithms critically miss out, in the human processing of music.