

The cognitive implausibility of statistical pattern recognition algorithms for music

Jean-Julien Aucouturier

Ikegami Laboratory, Graduate School of Arts and Science, Tokyo University, Japan

Keywords: Pattern Recognition, Signal Processing, Music Similarity, Saliency

We report on severe discrepancies between state-of-art computer models of music perception – based on the paradigm of statistical pattern recognition – and human performance. Typical models, proposed e.g. in the Music Information Retrieval Community, represent musical signals as the long-term statistical distribution of their local spectral features, a prototypical implementation of which being Gaussian Mixture Models of Mel-Frequency Cepstrum Coefficients. Such models make the implicit assumption that the perceptual importance of sound events correspond to their statistical predominance with respect to the global distribution. We study a typical instantiation of this approach, and compare its behaviour on two types of audio signals: urban soundscapes (e.g. sound recordings of a busy street) and polyphonic music (e.g. songs by *The Beatles*). We show that the saliency assumption holds very well for soundscapes, but not for polyphonic music (Fig. 1). The very informative frames for the modelling of the perception of polyphonic music (measured by their effect on the precision of the computer model) are the least statistically representative (the bottom 1%). Moreover, a large population of frames (in the range [60%, 95%]) is in fact detrimental to the modelling. In other words, music listeners routinely “hear” things that are not statistically significant in the actual signal, and that the low-level models of music perception studied here are intrinsically incapable of capturing. This is a strong case for investigating alternative models of how listeners assign saliency to sound events in a polyphonic music stream.

Figure 1. Comparison of the influence of the statistical importance of signal frames on the precision of a computer model of perceived audio similarity between sound textures, in the case of urban soundscapes and polyphonic music. The curve for music shows an unexpected non-monotonic behaviour, which notably shows that frames in statistical minority are in fact crucially important for the perception. This questions the underlying model of auditory saliency in state-of-art pattern recognition algorithms.

