Attention as Musical Interplay of Bottom-Up Accents and Expectation

Maarten Grachten, Ricard Marxer, Amaury Hazan, and Hendrik Purwins
Music Technology Group, Universitat Pompeu Fabra

Far from being passive and static, human perception is increasingly recognized as being an active process that involves expectation. We present a framework for modeling the human listening process in the context of melody perception. Grounded in empirical data, the framework has been adjusted to mimic attention allocation of newborns and adults with or without musical training. We focus on the modeling of tasks of higher order cognition and consequently use symbolic input streams. The setup consists of three layers: (1) feature detection, (2) saliency integration of bottom-up change detection and top-down expectation, and (3) allocation and strength calculation of attention.

The first layer consists in modularly organized perceptual feature detectors. This basic configuration permits qualitative validation against experimental data. The feature detectors provide symbolic streams of perceptual features localized in time, such as metrical emphasis, rhythm, melodic contour, pitch, tone center, and harmony. Two different representations of pitch and harmony are used to model different developmental stages and degrees of musical training.

In the second layer, the saliency of events in each channel (feature) is modeled. Saliency is determined by both bottom-up and top-down processes. The former is implemented as a quantization of a generalized first derivative of each channel output, the change strength (cf. Subfigure 2). The latter consists in forming and validating expectations. A Bayesian network is used as a simplified model for musical memory to make predictions. Cross-channel dependencies are modeled, facilitating the prediction of an event in one channel by events from the others. The network is updated incrementally to reflect observed event transitions. In this way, the network gradually learns to predict repeated patterns throughout the melody. The prediction errors (Subfigure 4 and 5) contribute to the saliency of events.

The third layer consists in the integration of the stimuli from the channels. This integration models the allocation of attention by a winner-takes-all approach. In our model, at each time it is indicated which perceptual feature determines the over-all attention. Salient events in an unattended channel attract the focus of attention. To avoid instability of the attentional process, changes of focus are followed by a refractory time.

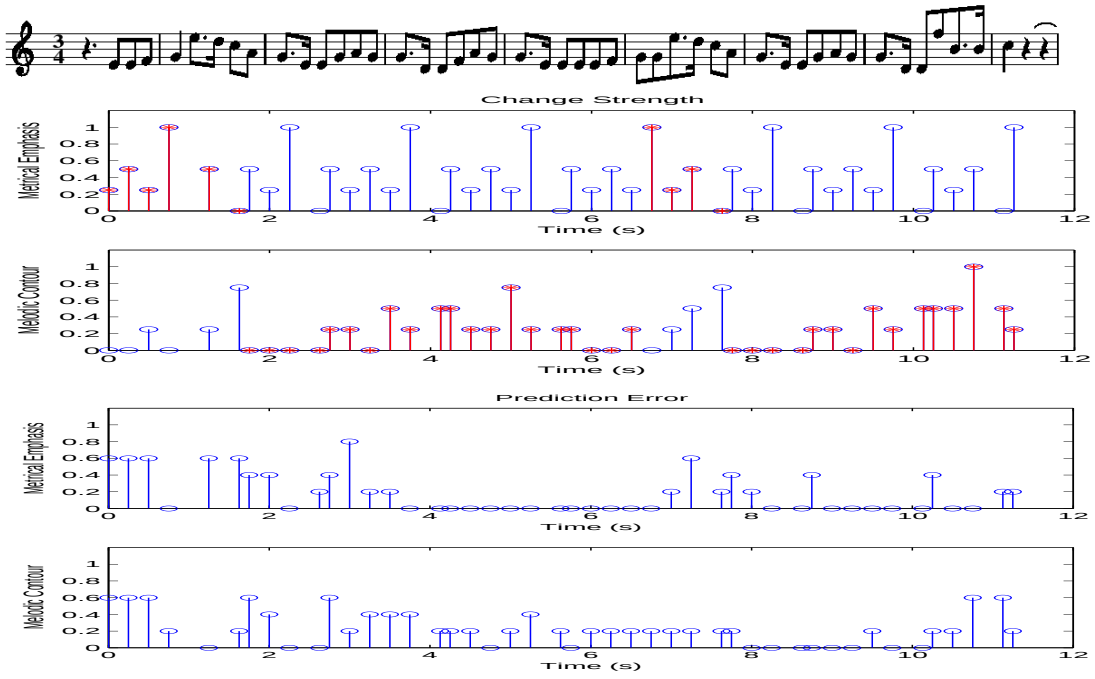

Figure 1: Attention to an Austrian folk song (1st row) is simulated in a listening model of a Western adult by bottom-up phenomenological accents (change strength, 2nd row), metrical accents (3rd row), attention focus (red filled circles), and top-down expectation (prediction error, 4th and 5th row).