How music listening programmes can be easily fooled

February 26 2015, by Bob L. Sturm

For well over two decades, researchers have sought to build music listening software that can address the deluge of music growing faster than our Spotify-spoilt appetites. From software that can tell you about the music you are hearing in a club and software that can recommend the music you didn't know you wanted to hear. From software that can intelligently accompany you practicing your instrument or act as an automated sound engineer, machine music listening is becoming

increasingly prevalent.

One particular area of intense research has been devoted to getting these "computer ears" to recognise generic attributes of [music](link): genres such as blues or disco, moods such as sad or happy, or rhythms like waltz or cha cha. The potential is enormous. Like spiders that crawl the web to make its information accessible to anyone, these machine listeners will be able to do the same for massive collections of music.

So far, we see some systems achieve accuracies that equal that of humans. The appearance of this human-level performance, however, may instead be a classic sign of unintentional cues from the experimentalist - a danger in experimental design recognised for over a century.

Clever Hans taps to an unexpected beat

My recent publication in the IEEE Transactions on Multimedia brings the cautionary tale of "Clever Hans" to machine listening research. Hans was a typical horse in Germany at the turn of the 20th century, with an atypical ability in abstract thought. Anyone could ask him to add several numbers, and away he would tap until he reached the correct answer. He could subtract, multiply, divide, factor, and even correctly answer questions written on a slate. His trainer also noted with pride that Hans learned to master new subjects with amazing ease, such as music theory, and the Gregorian calendar. However, give Hans an arithmetic problem to which no one knew the answer, or while he was blindfolded, and he was rendered back to the world of his more oat-minded barn mates.

It turns out that Hans really was clever, but only in the sense that he learned the most carrot-rewarding interpretation of the unconscious cues of the turn of the century German enquirer. It was apparently common to slightly bow one's torso while asking questions of a horse, and then

erect oneself at the moment it produced the correct answer. Hans had simply learned to begin tapping his hoof once he saw the torso bow, and stop once it unbowed, a cue so subtle that it eluded detection until properly controlled experiments were conducted. This brings us back to our music listening systems that appear to perform as well as humans.

Blindfolding and commissioning the listening machine

Taking one state of the art system measured to classify seven music rhythms with an 88% accuracy, we find that it only appears so because it has learned the most carrot-rewarding interpretation of the data it has seen: the generic rhythm labels are strongly correlated in the dataset with tempo. As long as a system can accurately estimate tempo, it can appear in this particular dataset to be capable of recognizing rhythm. If we slightly change the music tempi of the test dataset (like blindfolding Hans), our formerly fantastic system begins performing no better than chance.

To gain insight into what a state of the art [music genre](#) recognition system has learned, we have employed it as a kind of "professor" for young and naïve (computerised) music composition students. These students come to the professor with their random compositions, and it tells them whether or not their music is, for instance, unlike disco, quite like disco, or a perfect representation of disco. We keep the ones it says are perfect.

This system, which has been measured to recognise ten music genres with 82% accuracy (arguably human performance), confidently labelled 10 compositions as representative of each genre it has learnt (see the video below). By a listening experiment, we found that humans could not recognise any of the genres.

These problems in music listening systems are reflected in recent [studies of image content recognition systems](link), where small, insignificant changes to digital images can render a system unable to recognise something with which it had no problem previously.

We've used a similar procedure to make another state of the art [music listening](link) system label the same pieces of classical and country music as a variety of other genres. It is hard to hear much difference between them.

Just like horses, "horses" are not all bad

We should note that discovering a "horse" in an algorithm is not necessarily its ticket to the algorithmic glue factory. For one, a "horse" might be completely sufficient to meet the needs of a target application. It also provides an opportunity and mechanism to improve the validity of the system evaluation. And finally, discovering a "horse" provides a way to improve the system itself since it identifies the reasons for its behaviour. For that, I think even Clever Hans is clever enough to tap his hoof once for "Ja"!

  More information: "A closer look at deep learning neural networks with low-level spectral periodicity features":
[ieeexplore.ieee.org/xpl/articl … tp=&arnumber=6844511](link)

Provided by Queen Mary, University of London

provided for information purposes only.