Abstract

While music information retrieval (MIR) has made substantial progress in automatic analysis of audio similarity for Western music, it remains unclear whether these algorithms can be meaningfully applied to cross-cultural analyses of more diverse musics. Here we collect perceptual ratings from 62 participants using a global sample of 30 traditional songs, and compare these ratings against both pre-existing expert annotations and audio similarity algorithms. We find that different methods of perceptual ratings all produced similar, moderate levels of inter-rater agreement comparable to previous studies, but that agreement between human and automated methods is always low regardless of the specific methods used to calculate musical similarity. Our findings suggest that the MIR methods tested are unable to measure cross-cultural music similarity in perceptually meaningful ways.