Aditya Bhattacharjee

Emmanouil Benetos; Joren Six

Self-supervision in Audio Fingerprinting

The state-of-the art audio fingerprinting framework is not robust to pitch-shifting and time-stretching.

How to model the scalability of a search-retrieval task such as audio identification? Real-world performance of audio fingerprinting is measured at a scale which is orders of magnitude bigger than proposed experimental setups in the state-of-the-art. Is there a way to model the "capacity" of an embedding space?