Tutorial 5

Version Identification in the 20s

Furkan Yesiler, Christopher Tralie and Joan Serrà

Abstract

The version identification (VI) task concerns detecting and retrieving a set of songs that originate from the same underlying musical composition. Versions (or cover songs) convey the same musical entity while incorporating differences in several musical characteristics, including the differences in timbre, tempo, key, lyrics, and even added/deleted sections. The main applications include digital rights management and music catalog organization.

For more than a decade, VI systems suffered from the accuracy-scalability trade-off, with attempts to increase accuracy resulting in cumbersome, non-scalable systems. Recent years however have witnessed an increase in deep learning-based VI approaches that take a step toward bridging the accuracy-scalability gap, and we start seeing the possibility to deploy such systems in real-world applications. Although this trend positively influences the number of researchers and institutions working on VI, it may also result in obscuring the literature before the deep learning era. To appreciate the 20 years of novel ideas in VI and to facilitate building better systems in the next decade, we believe that now may be the right time to review some of the successful ideas and applications proposed in VI literature and connect them to current systems and ideas.

We will start the tutorial by explaining common input representations and feature post-processing steps. We will continue with comparing the pros and cons of alignment-based and embedding-based approaches, which constitute the two main perspectives for similarity estimation in VI. Lastly, after discussing a number of ideas that can be incorporated into any VI system, we will conclude by presenting the current challenges and future directions in VI research. Our goal is for the audience to leave with a thorough appreciation of both the history of the task and current directions, and that this context will allow them to jump into conducting novel research in the area.

Biographies

Furkan Yesiler is a PhD candidate at the Music Technology Group (MTG) of Universitat Pompeu Fabra (Barcelona). His research is focused on leveraging deep learning techniques to build accurate and scalable music version identification (VI) systems for industrial use cases. His recent contributions include MOVE, a state-of-the-art VI system based on musically-motivated principles; Da-TACOS, a large-scale VI benchmark set; and acoss, an open-source framework for feature extraction and benchmarking designed for VI. He received his MSc in Sound and Music Computing also from the MTG, with a focus on singing voice research. He graduated summa cum laude with two BSc degrees in computer engineering and industrial engineering from Koc University (Istanbul), where he was accepted with a full scholarship. During his bachelor's studies, he did internships in management consulting and M&A advisory companies in Istanbul, managed a student club with 200+ members, participated in a number of rowing competitions and musical theater shows, and spent a trimester at the University of California, Santa Barbara.

Christopher Tralie is an assistant professor in Math and Computer Science at Ursinus College in Collegeville, Pennsylvania, USA. He works in applied geometry/topology and geometric signal processing, and his work spans shape-based music structure analysis and version identification, video analysis, multimodal time series analysis, and geometry-aided data visualization. He received a B.S.E. from Princeton University 2011, a master's at Duke University in 2013, and a Ph.D. at Duke University in 2017, all in Electrical Engineering. His Ph.D. was primarily supported by an NSF Graduate Fellowship, and his dissertation is entitled "Geometric Multimedia Time Series." He did a postdoc at Duke University in Mathematics and a postdoc at Johns Hopkins University in Complex Systems. He was awarded a Bass Instructional Teaching fellowship at Duke University, and he maintains an active interest in pedagogy and outreach, including longitudinal mentoring of underprivileged youths in STEAM (STEM + arts) education.

Joan Serrà is a staff researcher with Dolby Labs in Barcelona since 2019, where he works on deep learning and audio processing. He did his MSc (2007) and PhD (2011) in automatic version identification at the Music Technology Group of Universitat Pompeu Fabra. He also did a postdoc in artificial intelligence at IIIA-CSIC (2011-2015). After that, he was a machine learning researcher at Telefónica R&D (2015-2019). He has had research stays at the Max Planck Institute for the Physics of Complex Systems (2010) and the Max Planck Institute for Computer Science (2011). He has been involved in more than 10 research projects, funded by National and European institutions, and co-authored over 100 publications, many of them highly-cited and in top-tier venues, including NeurIPS, ICLR, ICML, InterSpeech, ICASSP, and ISMIR.