OPTIMIZING DTW-BASED AUDIO-TO-MIDI ALIGNMENT AND MATCHING

Colin Raffel and Daniel P. W. Ellis

LabROSA
Columbia University
New York, NY

ABSTRACT

Dynamic Time Warping (DTW) has proven to be an extremely effective method for both aligning and matching recordings of songs to corresponding MIDI transcriptions. The performance of DTW-based approaches in this domain is heavily effected by system design choices, such as the representation used for the audio and MIDI data and DTW's adjustable hyperparameters. We propose a method for optimizing the design of DTW-based alignment and matching systems. Our technique uses Bayesian optimization to tune system design and hyperparameters over a synthetically created dataset of audio and MIDI pairs. We then perform an exhaustive search over DTW score normalization techniques in order to determine an optimal method for reporting a reliable alignment confidence score, which is necessary for matching tasks. Using our approach, we are able to create a DTW-based system which is conceptually simple and highly accurate at both alignment and matching. We also verified that our system achieves high performance in a large-scale qualitative evaluation of results on real-world data.

Index Terms— Dynamic Time Warping, Audio to MIDI Alignment, Sequence Retrieval, Bayesian Optimization, Hyperparameter Optimization

1. INTRODUCTION

MIDI files can provide a bounty of ground-truth data for content-based music information retrieval (MIR) tasks, including beat/bar tracking, onset detection, key estimation, automatic transcription, and score-informed source separation [1, 2, 3, 4]. However, in order to utilize this data, a given MIDI file must first be aligned in time to a recording of the piece it is a transcription of. A related problem is MIDI-to-audio matching, where we are given separate collections of MIDI and audio files and must determine which audio file (if any) each MIDI file corresponds to. This problem is motivated by the dearth of metadata information typically available in MIDI files, making any kind of text-based matching infeasible [5]. With or without useful metadata, it is beneficial to be able to produce a confidence score for alignment which communicates how well the content in a MIDI file matches a given audio file, which can help determine the transcription quality.

Despite these tasks' complementary nature, most previous research has focused on systems meant for either alignment or matching. In the context of MIDI-to-audio alignment, a wide variety of techniques have been used, which typically involve determining a correspondence between discrete times in the audio and MIDI file. While approaches inspired by edit distance metrics such as Smith-Waterman [1] and Needleman-Wunsch [6] have been used, arguably the most common approach is Dynamic Time Warping (DTW) [7]. First proposed for comparing speech utterances [8], DTW uses dynamic programming to find a monotonic alignment such that the total distance between aligned feature vectors is minimized. This property makes it well-suited for audio-to-MIDI alignment when we expect that the MIDI is an accurate continuous transcription (i.e. without out-of-sequence or incorrect sections).

A nice property of DTW is its appropriateness as a measure of the "similarity" between two sequences. This is thanks to the fact that the total distance between pairs of feature vectors in the DTW-based alignment can be used reliability as a distance metric. In fact, DTW has seen extensive use solely as a way of measuring sequence similarity in the data mining literature [9]. In the context of MIDI and audio files, [10] evaluated the effectiveness DTW distance to match a small collection of Beatles MIDIs to recordings of Beatles songs.

Despite its widespread use, DTW's success can be highly dependent on its exact formulation as well as system design choices such as the feature representation and local distance metric used. To our knowledge, there has been no large-scale quantitative comparison of different DTW-based alignment systems. This is likely due to the fact that evaluating its performance would require either a collection of MIDI and audio pairs for which the correct alignment is already known (which does not exist) or manual audition and rating of the output of the system (which is time-consuming). The present work aims to remedy this and produce a DTW-based alignment system which is optimal in terms of both alignment and matching accuracy.

After giving an overview of typical DTW-based alignment systems (Section 2), we propose a method for creating a synthetic dataset of MIDI-audio pairs by applying realistic corruptions to MIDI files, which allows us to know a priori the correct alignment (Section 3). We then tune hyperparameters for alignment using Bayesian optimization (Section 4) and for confidence reporting using an exhaustive search (Section 5). Finally, we perform a large-scale qualitative evaluation of our proposed alignment system on real-world data (Section 6) and discuss possible avenues for improvement (Section 7).

2. DTW-BASED ALIGNMENT

Formal definition of DTW-based alignment, with parameter and representation discussion

Discussion of extracting a confidence score, normalization methods

3. CREATING A SYNTHETIC ALIGNMENT DATASET

Collection of MIDI files

"Easy" corruption, for alignment accuracy

"Hard" corruption, for matching accuracy

4. OPTIMIZING DTW-BASED ALIGNMENT

Short overview of Bayesian optimization

Parameter space (including multiplicative penalty)

Random trials

Discussion of best-performing aligners; also best aligner with beats

5. OPTIMIZING CONFIDENCE REPORTING

Grid search

Statistical tests used

Choosing the best alignment scheme (with algorithm box?)

6. QUALITATIVE EVALUATION

Data preparation

Evaluation criteria

Results

7. AVENUES FOR IMPROVEMENT

Augmentation with MUDA, partial alignments, robustness to missing instruments, re-training specifically on subsequences

8. REFERENCES

[1] Sebastian Ewert, Meinard Müller, Verena Konz, Daniel Müllensiefen, and Geraint A. Wiggins, "Towards cross-version harmonic analysis of music," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 770–782, 2012.

[2] Robert J. Turetsky and Daniel P. W. Ellis, "Ground-truth transcriptions of real music from force-aligned MIDI syntheses," *Proceedings of the 4th International Society for Music Information Retrieval Conference*, pp. 135–141, 2003.

[3] Sebastian Ewert, Bryan Pardo, Mathias Muller, and Mark D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[4] Colin Raffel and Daniel P. W. Ellis, "Intuitive analysis, creation and manipulation of MIDI data with `pretty_midi`," in *Proceedings of the 15th International Society for Music Information Retrieval Conference Late Breaking and Demo Papers*, 2014.

[5] Colin Raffel and Daniel P. W. Ellis, "Large-scale content-based matching of MIDI and audio files," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, 2015.

[6] Maarten Grachten, Martin Gasser, Andreas Arzt, and Gerhard Widmer, "Automatic alignment of music performances with structural differences.," in *Proceedings of the 14th International Society for Music Information Retrieval Conference*, 2013, pp. 607–612.

[7] Meinard Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[8] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.

[9] Donald J Berndt and James Clifford, "Using dynamic time warping to find patterns in time series.," in *AAAI Workshop on Knowledge Discovery in Databases*, 1994, vol. 10, pp. 359–370.

[10] Ning Hu, Roger B. Dannenberg, and George Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 185–188.