Bottlenecks and Solutions for Audio to Score Alignment Research

Alia Morsi, Xavier Serra

MTG, Universitat Pompeu Fabra

Abstract

Identifying the current SOTA for Audio to Score Alignment (ASA) is not straightforward.

This is due to the variation across prior work with respect to system choices, data, and evaluation.

This hinders progress in ASA research since it becomes harder to determine concrete improvement directions.

Paper Goals

- Demonstrate the extent of variation across prior work.
- Propose solutions (both conceptual and practical)
- Provide an example of expanding the usable data, taking a step forward in one of our proposed solutions.

Sources of Variation

1. Scope/use-case differences

A variety of practical scenarios have been addressed across prior work. Examples include:

- **Different Instruments** (Piano, Multi-instrument, Clarinet).
- **Deviations in score quality.**
- **Presence of performance mistakes.**

Often, each scenario is treated as a different alignment problem, and evaluated with different data.

An ASA system itself includes choices tied to its primary use-case, and therefore is only evaluated with metrics and data relevant to that use-case.

Hence, many systems are proposed but are not compared in a unified way.

2. Different Alignment Modalities

Sequence comparisons are sometimes performed in the **audio modality**, in the **symbolic modality**, or with **intermediate representations** between Audios and Scores/Score Images.

3. Differences in Approach

Dynamic Time Warping (DTW) and **Hidden Markov Models (HMM)** are quite popular.

There is lots of variation cross ASA DTW systems.

4. Data Differences

Datasets include: The Vienna 4x22 Piano Corpus, Bach10, RWC Database, MUS Subset of MAPS Database, private data, synthetic data, ..

5. Evaluation Differences

Although most metrics revolve around the rates of Alignment (or Misalignment), and the Alignment Error, they are not very standardized.

Suggested Solutions

- Define ASA uniquely with the **scope of musical scenarios representing its core,** and treat ASA as one multi-faceted problem, rather than several separate problems.
- Establish a **benchmarking framework** with metrics and relevant data covering all such scenarios.
- Each proposed ASA system would be evaluated for all the agreed upon practical scenarios.
- This is only possible by
  - Consolidating existing relevant datasets.
  - Expanding the usable data in both size and coverage of scenarios, by reusing candidate datasets, and with synthesis.

We do not undermine the impact of a target use-case on system choices.

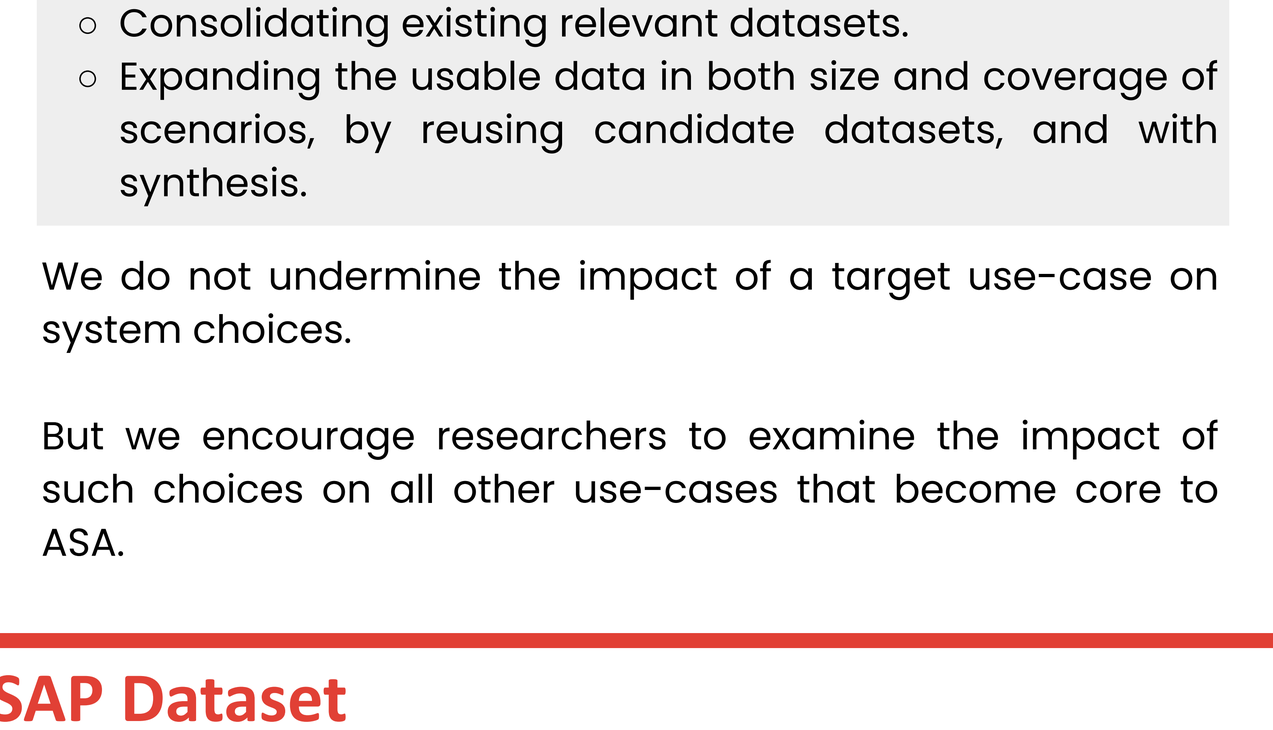
But we encourage researchers to examine the impact of such choices on all other use-cases that become core to ASA.

Data Extension Example: Reusing the ASAP Dataset

- Candidate datasets are those which have a level of alignment between audio and scores.
- We re-use the Aligned Scores and Performances (ASAP) dataset, which has 520 beat aligned score-performance pairs.
- Since the performances are monotonic, we use Piecewise Linear Interpolation to approximate alignments between beats.

Potential Data Problems, and Filtering

- Misalignments in approximated alignment ground truths due to low resolution (temporal sparsity of beats),
- Potential Misalignments in original Dataset



Conclusions

Extending the definition of ASA must be accompanied by extending the available data to cover a variety of ASA use-cases.

This data extension example for ASAP alone is far from sufficient to help ASA research get past its current bottleneck.

It is a starting point, **and an opportunity to discuss more solutions with the research community.**

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN

upf. MTG Music Technology Group