Evaluation of Real-Time Audio-to-Score Alignment

Arshia Cont, Diemo Schwarz, Norbert Schnell, Christopher Raphael

EVALUATION OF REAL-TIME AUDIO-TO-SCORE ALIGNMENT

Arshia Cont
Ircam UMR CNRS 9912
CRCA, UCSD
cont@ircam.fr

Diemo Schwarz, Norbert Schnell
Ircam–Centre Pompidou, Paris
UMR CNRS 9912
{schwarz, schnell}@ircam.fr

Christopher Raphael
Indiana University
Bloomington, IN
craphael@indiana.edu

ABSTRACT

This article explains evaluation methods for real-time audio to score alignment, or *score following*, that allow for the quantitative assessment of the robustness and preciseness of an algorithm. The published ground truth data base and the evaluation framework, including file formats for the score and the reference alignments, are presented. The work, started for MIREX 2006, is meant as a first step towards a standardized evaluation process contributing to the exchange and progress in this field.

1 INTRODUCTION

Score following is the real-time alignment of a known musical score to the audio signal produced by a musician playing this score, usually in order to synchronise an electronic accompaniment of the music to the performer, leaving him with all possibilities of expressive performance. Despite its history of more than 20 years [1], only very few attempts of a systematic evaluation of the result of score following have been made, and therefore, representation of research work is usually limited to demonstration of short examples or *subjective* ways to assess the quality of such systems.

Evaluation gives an indication of the quality of an alignment algorithm and allows the comparison of different methods, implementations, parameters, etc., and the quantification of improvements gained by training. The lack of a unified common evaluation framework has led to different and personal approaches to the problem of score alignment. This clearly makes further progress of the problem more difficult and assessment of the results uncertain.

2 EVALUATION METHOD

We start by defining five basic *event measures* from which the *assessment metrics* are calculated:

The **error** $e_i = t_i^e - t_i^r$ is defined as the time lapse between the alignment positions of corresponding events in the reference t_i^r and the estimated alignment time t_i^e for score events i. A real-time system cannot correctly detect a note until sometime after it has occurred. We define the **latency** $\ell_i = t_i^d - t_i^e > 0$ of a detection to be the difference between the time a dectection is made t_i^d and the estimated note onset time. The **offset** $o_i = t_i^d - t_i^r$ is the lag between the time the event occurred t_i^r and the reporting of the detection t_i^d, which is important for purely reactive systems.

Missed notes are events that are not recognized, i.e. that exist in the reference but are not reported. **Misaligned notes** are events in the score that are recognized but are too far (regarding a given threshold θ_e, e.g. 300 ms) from the reference alignment to be considered *correct*.

In fact, due to score–performance mismatch and misses or false matches, the events recognised by the alignment algorithm do not necessarily correspond one-to-one to the reference events. For instance, if a recognizer is uncertain about the onset location of a note, it may be better not to report that onset rather than report it incorrectly. Thus, when evaluating a score follower, it makes sense to distinguish between missed notes and misaligned notes.

Given these event measures, the *assessment metrics* characterizing the quality of an alignment are then: The **miss rate** p_m is the percentage of missed score events. The **misalign rate** p_e is the percentage of misaligned events with their absolute error $|e_i|$ greater than θ_e. They constitute cases in which the recognizer is not merely inaccurate, but simply mistaken. **Piece completion** p_c is the percentage of the events that was followed until the follower hangs, i.e. from where on there are only misaligned events. The **average latency** μ_ℓ for non-misaligned events, is an overally measure of the latency of the system. The **average absolute offset** μ_o for non-misaligned events gives an indication of the reactivity of the follower. The **variance of error** σ_e is the standard deviation of e_i for non-misaligned events and shows the imprecision or spread of the alignment error. The **average imprecision** μ_e is the average absolute error of non-misaligned events and shows the global imprecision.

Systems are evaluated by two measures: *Piecewise precision rate* as the average of the percentage of correctly detected notes for each piece group in Table 1, and *overall precision rate* on the whole database.

3 EVALUATION FRAMEWORK

The evaluation procedure is as follows: Each score following system to be evaluated has to implement the *evaluation interface* that allows for the insertion of the system into the evaluation procedure. It defines the way the system is invoked, the order of the parameters passed to the system, and the way the results are returned. It is a simple commandline call with as parameters the names of the score and audio files to be read, and the path to the output file to be written. In general some glue code must be written to interface the system with the evaluation framework. The *batch processor* invokes and controls the score following system using the interface with each pair of score

and performance from the evaluation database. The *evaluator* then reads all alignment results, and calculates and outputs statistics of the evaluation measures described in section 2.

The *file format definitions* pertain to the performance audio files, the score files, the reference alignment files, and the alignment result files: For the **audio files**, the AIFF and RIFF-wave standards are sufficient. The **score files** pose the well known problem of a missing standard for a score representation format that fulfills all our requirements. In 2006, we used MIDI files, but already such a simple musical event like a trill could not be adequately represented in MIDI. Therefore, we defined a text format [1] with one line per event and a fixed number of columns. One column carries a unique ID for each event, valid across all file types, others the score position in metric and time position t_i^s. The **reference alignment** files constitute a ground truth alignment between a score and its performance. They reuse the score file format, replacing the score time with the reference alignment time t_i^r. The two types of **result files** represent the alignment found by a score following system between a score and a recording of a performance of it. It is again based on the score file format, with first the estimated event onset time in t_i^e and second the reporting time relative to the performance audio file t_i^d at the place of t_i^s.

4 GROUND TRUTH DATABASE

The database of ground truth alignments is the most valuable asset of the evaluation framework, and the most tedious to assemble. Since the performance of a particular score following system usually depends on the instrument and the repertoire (i.e. style) of the chosen pieces it is desirable that the reference database covers a wide range of instruments and styles. As the usual pragmatic compromise, the database is constituted by contributions from the participants in the evaluation exchange.

The current database, detailed in Table 1, has been constituted by the authors and used for the first evaluation exchange in 2006. There are various solo instruments in the database and we have classical music as well as contemporary music performances. An example from the database is available on the web. [2]

Piece name	Composer	Instrument	Files	Duration	Events
Explosante-Fixe	Boulez	Flute	47	17:10	2022
Violin Sonatas	Bach	Violin	3	13:50	3996
K. 370	Mozart	Clarinet	4	14:44	2710
Dorabella	Mozart	Voice	4	01:44	229
Total			58	47:38	8957

Table 1. MIREX06 Score Following Reference Database

In order to achieve high-resolution alignment, we first gathered a database of monophonic (or slightly polyphonic) audio with their score, and reference alignments that give the time-position of each note in the recordings.

The references can come from two sources: For simple pieces to align, the output of a score following system itself, is precise enough to be used as reference, after format conversion, verification, and manual correction if necessary (see below). Although this invalidates the data to be used for comparison of different systems because it would confer an unfair advantage to the system generating the data, it is useful to do regression testing of a follower, and to measure improvements by training or changes in the code. For more complex pieces, an existing off-line audio-to-score alignment system based on dynamic time-warping (DTW) [2] is used. However, even after off-line DTW alignment, the preciseness of the segmentation is sometimes not sufficient. For this reason, methodologies for hand correction of the alignment results were developed, that combine the off-line alignment with transient marks found by an onset detection algorithim, overlaid over the spectrogram.

5 CONCLUSION AND FUTURE WORK

Our evaluation methodology and framework allow for the first time to assess the robustness and preciseness of a score following system in a way that gives objective and repeatable results. This enables us to track improvements in the code, parameters, or training data sets of a score following system, and to compare different systems.

This methodology has also been applied to a first MIREX evaluation exchange amongst research groups with related interests, and it will be improved and its applicability enlarged to allow more systems to be evaluated and compared this year. Results of the evaluation process defined in this paper are available through the MIREX 2006 result page. [3] The presented evaluation methodology could possibly be extended to other variants of alignment such as MIDI-to-score alignment, audio-to-audio alignment, and off-line score-to-audio alignment.

6 ACKNOWLEDGEMENTS

We would like to thank Miller Puckette, Roger Dannenberg for sharing their experience and contributing recordings and scores to the evaluation database, and all participants in the MIREX score following evaluation endeavour. This work is partially conducted in the framework of the i-Maestro project, partially supported by the European Community under the IST Information Society Technologies priority of the 6th Framework Programme for R&D. [4]

7 REFERENCES

[1] Nicola Orio, Serge Lemouton, Diemo Schwarz, and Norbert Schnell. Score Following: State of the Art and New Developments. In *New Interfaces for Musical Expression (NIME)*, Montreal, 2003.

[2] Ferrol Soulez, Xavier Rodet, and Diemo Schwarz. Improving Polyphonic and Poly-Instrumental Music to Score Alignment. In *ISMIR*, Baltimore, 2003.

[1] http://www.music-ir.org/mirex2007/index.php/Score_File_Format

[2] http://crca.ucsd.edu/arshia/mirex06-scofo/

[3] http://www.music-ir.org/mirex2006/index.php/Score_Following_Results

[4] IST-026883 http://www.i-maestro.org