

Searching Musical Audio Using Symbolic Queries

Based on *Cross-Domain Content-Based Retrieval of Audio Music
through Transcription*, PhD Thesis, RMIT

Iman S. H. Suyoto
ade@ishs.net

10 March 2010

Fakultas Ilmu Komputer Universitas Indonesia

Problem

- A content-based music retrieval system typically works with either raw audio (e.g. MP3, WAV, AU) or symbolic sequences (e.g. MIDI, MusicXML, LilyPond).
- An audio query is matched with a collection of audio whereas a symbolic query sequence is matched with a collection of symbolic sequences.
- Some previous work:
 - Audio: Logan & Salomon (2001), Aucouturier & Pachet (2002), Pampalk et al. (2005), Somerville & Uitdenbogerd (2005),
 - Symbolic: Uitdenbogerd & Zobel (1999), Suyoto & Uitdenbogerd (2004, 2005, 2006), Grachten et al. (2005), Orio (2005),

Research Question

Is it possible to retrieve audio by its symbolic equivalent representation (whether in full or partial)?
(This task can also be called cross-domain content-based retrieval)

Cross-domain Retrieval

Overview

Background
Research
question
Background

Methodology

Transcription
Standardisation
Alignment
Full queries
Short queries

Discussion

Conclusions

Ending

- Cross-domain content-based retrieval is closely related to score-to-audio alignment, e.g. Müller et al. (ISMIR 2004), Raphael (ISMIR 2004), Shalev-Shwartz et al. (ISMIR 2004), and Soulez et al. (ISMIR 2003).
- Score-to-audio alignment:
 - Typically has its effectiveness measured in a function of number of correct or incorrectly identified notes.
 - Is concerned with structural similarity contained within a piece.
- Retrieval:
 - Is more concerned with the ability to discriminate correct answers from incorrect ones and ranking the correct ones highly.
 - Requires approximate matches to be ranked highly.

Prior Work

- Pickens et al. (ISMIR 2002): full-query task, polyphonic symbolic collection, polyphonic audio queries. MAP = 0.479.
- Hu et al. (WASPAA 2003): full-query task, polyphonic symbolic collection, polyphonic audio queries; all the Beatles' songs. Mean precision at 1 = 0.49.
- Shalev-Shwartz et al. (SIGIR 2002): short-query task, polyphonic audio collection, monophonic symbolic queries. Average precision = 95%.
- Marolt (ISMIR 2006): full-query task, polyphonic audio collection, polyphonic audio queries. 27% of his in the top 5 returned answers.

Matching

Matching phases:

- 1 Transcription.
- 2 Standardisation.
- 3 Alignment.

Transcription

- Audio files in collection transcribed to symbolic data.
- Transcriber: TS-AudioToMidi 3.30.
- Transcription results saved as MIDI files.
- So, now just match symbolic queries with the transcription results. That simple?
- Problem: transcription result contains noise. Example:
Actual performance: (J. S. Bach's "BWV 1007 Prelude" performed by Carrai)



is transcribed into ...

Problem with Transcription

Searching
Musical Audio
Using
Symbolic
Queries

Suyoto

Overview

Background
Research
question
Background

Methodology

Transcription
Standardisation
Alignment
Full queries
Short queries

Discussion

Conclusions

Ending

The image displays a musical score consisting of four systems of staves. Each system includes a bass clef staff on the left and a treble clef staff on the right. The music is written in 4/4 time and is characterized by a high density of notes and accidentals (sharps and naturals), which makes it difficult to transcribe accurately. The notation is complex, with many notes beamed together and numerous accidentals throughout the piece.

Problem with Transcription

Overview

Background
Research
question
Background

Methodology

Transcription
Standardisation
Alignment
Full queries
Short queries

Discussion

Conclusions

Ending

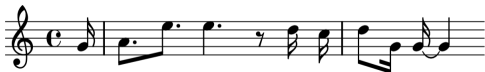
- Contains too much noise.
- Caused by various factors: reverb, timbre characteristics, performer's friction with the instrument,
- A noise removal procedure was previously thought to be needed, but actually it's not needed.

Standardisation

- *Standardisation*: technique of generation of a string that represents a melody.
- The idea: approximate string matching between a query and the tunes in collection.
- Relative-pitch standardisations have proven to be highly effective for symbolic retrieval, but not for our task.
- We use *pitch classes* instead.

Standardisation

- A note is represented by its pitch class, i.e. the pitch name only, its octave is stripped. Example: "G A E E D C D G G".



- Chords are "flattened." Example: "B \flat D F B \flat D F B \flat D F B \flat D F B \flat C E \flat G B \flat C E \flat G B \flat C E \flat B \flat D F".



- Disadvantage: doesn't support transposition-invariant matching, e.g. the C4-D4-E4 and G4-A4-B4 don't match. However, this supports much higher retrieval effectiveness compared to relative-pitch standardisation.

Alignment

- Full-query task: Can a manually constructed symbolic sequence be matched with a transcription-of-audio sequence, both of which are the same tune (albeit different renditions)?
- Short-query task: Can a short query be used to retrieve whole tunes?

Full-Query Matching

- The well-known *longest common subsequence* (LCS) algorithm is used to align a query q with an audio transcription a .
- Musician: The query is transposed to all possible keys.
CS geek: The query string is transposed across the alphabet C C# D D# E F F# G G# A A# B.
- Align the transposed queries. The highest obtained LCS score is normalised:

$$S_y(q, a, y) = \text{LCS}(q, a) \ln^{-y} |a|$$

- For our collection, the best $y = 2.0$.

Experimental Setup

Collection:

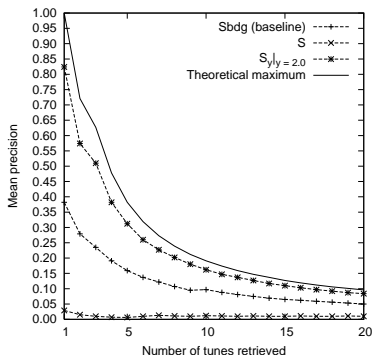
- Magnatune classical music collection (as on 28 April 2005) stored as MP3 files (stereo).
- Contains cover versions of the same repertoires.
- Transcribed using TS-AudioToMidi (default settings). 1 808 MIDI files obtained.

Queries:

- MIDI versions of some of the covers in Magnatune.
- Sources: the Mutopia Project, Kern Scores, and MuseData.
- 34 queries with at least a relevant answer.

Baseline: SBDG in Suyoto & Uitdenbogerd (RMIT TR-07-1):
<http://mirt.cs.rmit.edu.au/pubs/sbdg/>.

Results



We tried varying $y \in \{1.0, 1.1, 1.2, \dots, 3.0\}$. $y = 2.0$ gives the best MAP = 0.826. $S_y(q, a, 2.0)$ is significantly more effective than SBDG. LCS scores without normalisation don't have sufficient discriminatory power.

Short-Query Matching

Overview

Background
Research
question
Background

Methodology

Transcription
Standardisation
Alignment
Full queries
Short queries

Discussion

Conclusions

Ending

- A sliding window is used. Window size: $W + 1$.
 $W = \lceil 2d|q| \rceil$. d is a tunable window size parameter.
- For $z \leftarrow a_0 \dots a_W$, $s = \text{LCS}(q, z)$.
- Slide by $\lceil d \rceil$. $z \leftarrow a_{\lceil d \rceil} \dots a_{\lceil d \rceil + W}$. $s' = \text{LCS}(q, z)$.
 $(s' > s) \rightarrow (s \leftarrow s')$.
- Slide by $\lceil d \rceil$. $z \leftarrow a_{2\lceil d \rceil} \dots a_{2\lceil d \rceil + W}$. $s' = \text{LCS}(q, z)$.
 $(s' > s) \rightarrow (s \leftarrow s')$.
- Repeat while $n\lceil d \rceil + W < |a|$, $n \geq 0$.
- Repeat for all transpositions. The similarity score between q and a is $S_d(q, a, d) =$ the final value of s .
- Anyone who wants to collaborate to improve the efficiency of this algorithm?

Results

- The same collection and original queries are used.
Baseline: SBDL100.
- We tried varying $|q| \in \{20, 50, 100\}$ ($d = 1$). Best at $|q| = 100$: MAP = 0.745, lower than S_y , but P@1 is the same. The first $|q|$ symbols are used.
- We tried varying $d \in \{1, 2, 3, 4, 5\}$ ($|q| = 100$). Best at $d = 1$.
- We tried fine-tuning $d \in \{0.5, 0.6, 0.7, \dots, 1.5\}$. Best at $d = 1.1$: MAP = 0.768.
- We tried varying the locations of query substrings: beginning, middle, and ending. MAPs: 0.768, 0.651, 0.566. Still better than the baselines SBDG (MAP: 0.374) and SBDL100 (MAP: 0.229).

Discussion

- Average full query length = 814.15. Only 12.3% of this is needed for effective retrieval.
- Greater d increases the probability that a query obtains the maximum possible score, but also leads to many incorrect answers being scored highly. Smaller d has better discriminatory power.
- The LCS algorithm is appropriate as the amount of noise between matching symbols is very high, thereby lowering the effectiveness of applying penalties for non-match operations.
- The transcription process inserts a lot of noise, rendering an interval-based representation unreliable.

Conclusions

- Effective retrieval of audio by using symbolic queries is possible, with MAP of 77%.
- Truncated queries are as effective as full-length ones. Moreover, they reduce the burden on users by allowing them to issue short queries.
- Truncated queries are effective no matter whether they are from the beginning, middle, or end of a piece.
- Representing notes as their pitch classes facilitates more effective retrieval than using a relative pitch standardisation, because the former is not severely affected by the noise in transcriptions.

More Information

The thesis contains results with larger testbeds. The methods still work well with those testbeds.

The audience is encouraged to read the thesis for more information: <http://tr.im/JitM>.

Questions?



Source: Flintlocke's Guide to Azeroth, Episode 13, Part 12.