Adaptation of String Matching Algorithms for Identification of Near-Duplicate Music Documents

M. Robine, P. Hanna, P. Ferraro and J. Allali

University of Bordeaux 1, LaBRI, SCRIME

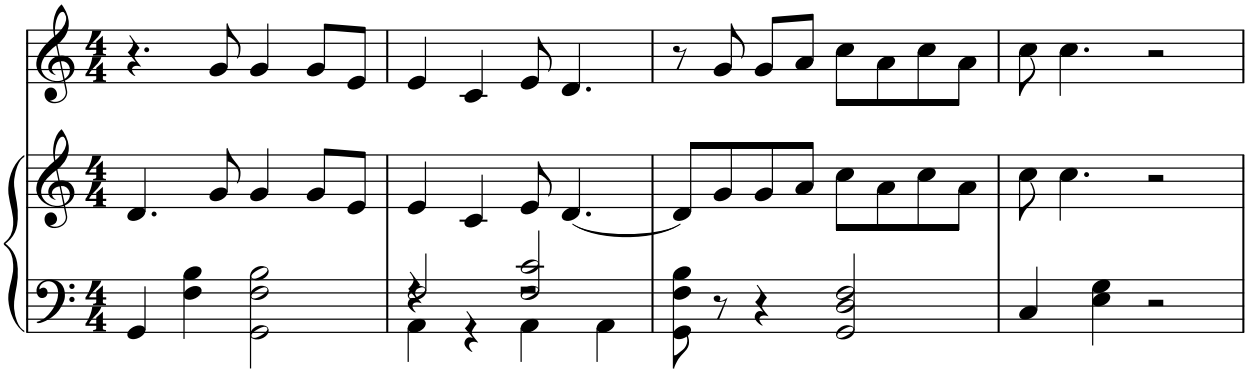
Introduction

- Motivation: Plagiarism detection
- What is considered as plagiarism in music ?
- Estimation of the similarity of music documents
- Huge music databases
- Symbolic representation of music

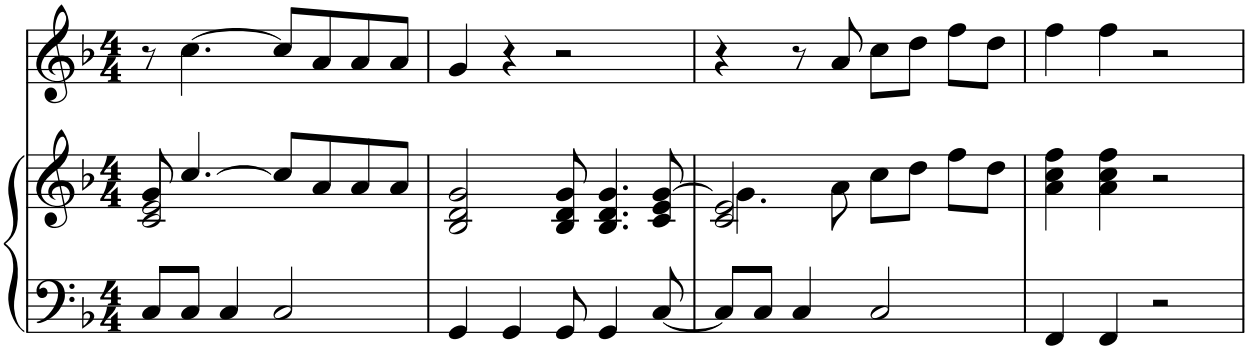
Bright Tunes Music v. Harrisongs Music (1976)

He’s So Fine (Ronald Mack) / My Sweet Lord (George Harrison)

He's So Fine



My Sweet Lord



Similarity between sequences

→ String matching algorithms

Edit operations:

- Insertion (I)
- Deletion (D)
- Matching (M)
- Substitution (S)

Example:
distance(`APPLIED`,`PRINCE`) ?

string 1	A	P	P	L	I	_	_	E	D
string 2	_	P	R	_	I	N	C	E	_
operation	D	M	S	D	M	I	I	M	D

Edit-Distance : local alignment

[Smith et al. 1981]

- Based on dynamic programming
- Determines the region of best match for two sequences
- Outputs:
  - How good the best local alignment is $\Longrightarrow$ **score**
  - Positions corresponding to this best local alignment
- Each operation is associated to a score (may depend on the values of the sequences compared). For example:
  - Deletion/Insertion: -2
  - Substitution: -1
  - Matching: 1
- No negative score

Local alignment

	_	P	R	I	N	C	E	S	S
_	0	0	0	0	0	0	0	0	0
R	0								
I	0								
C	0								
E	0								

$$M[i,j] = \max\begin{cases} 0 \\ M[i-1,j]-2 \\ M[i,j-1]-2 \\ M[i-1,j-1]+\text{match}(\text{string1}[i],\text{string2}[j]) \end{cases}$$

Local alignment

	_	P	R	I	N	C	E	S	S
_	0	0	0	0	0	0	0	0	0
R	0	0	1	0	0	0	0	0	0
I	0	0	0	**2**	0	0	0	0	0
C	0	0	0	0	1	1	0	0	0
E	0	0	0	0	0	0	**2**	0	0

$\Longrightarrow$ Similarity score = 2

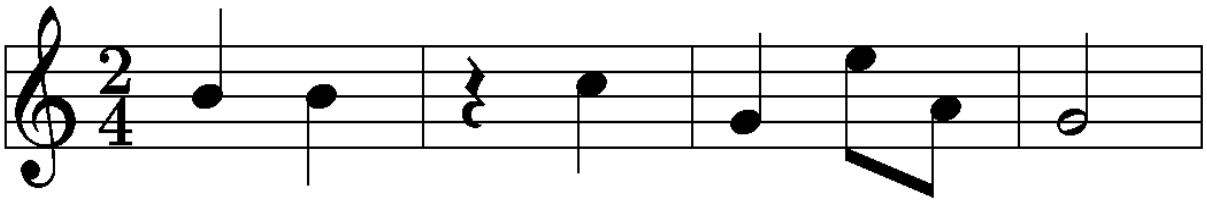
corresponding to the alignment:

P	R	I	N	C	E	S	S
_	R	I	_	C	E	_	_

Representations of music

Each note can be represented by a **pitch** and a **duration** [Mongeau et al. 1990]

Example:



represented by the sequence:

```
(B4 B4 r4 C4 G4 E2 A2 G8)
```

Improvements for Music Documents

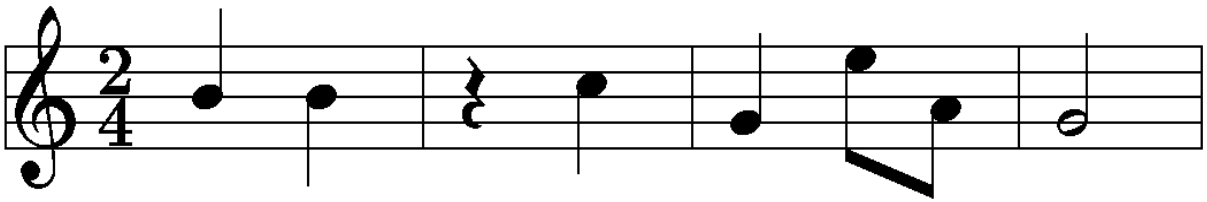
- Tempo invariant
- Transposition invariant
- Representation of polyphony

Weighting options:

- Consonant interval for substitution [Ferraro et al. 2007]
- Music theory elements [Robine et al. 2007]

Transposition invariant

Sequence of intervals instead of pitches



- Exact interval: number of half-tones

  $$0, 1, 5, 9, 7, 2$$

- Modulo 12 interval

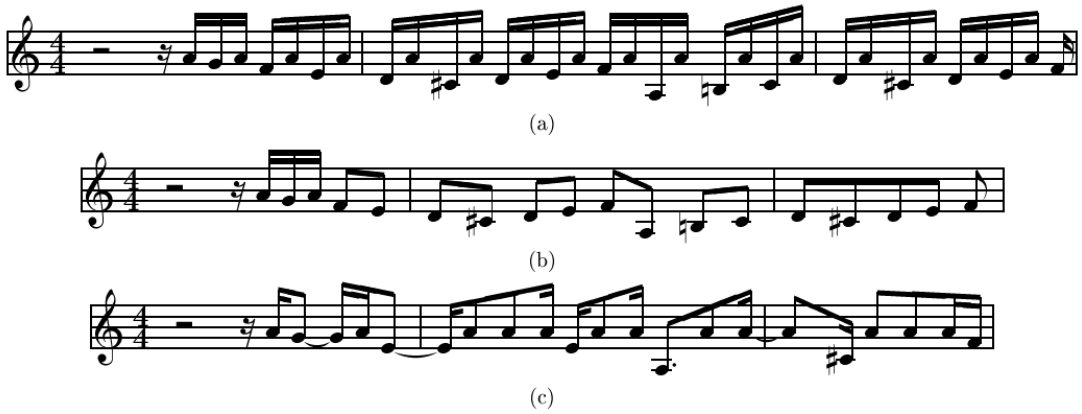
  $$0, 1, 5, 3, 5, 2$$

- Directed modulo 12 interval

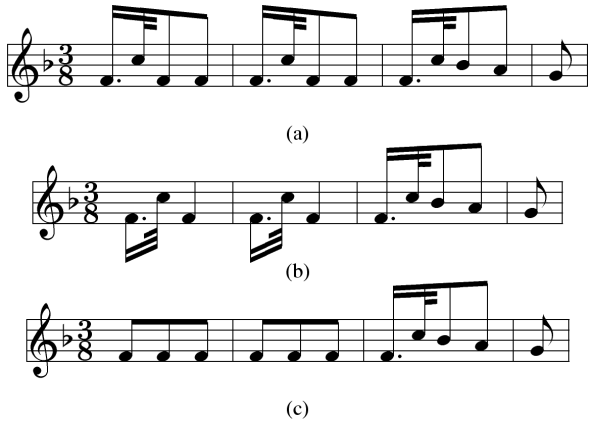
  $$0, +1, -5, +3, -5, -2$$

Music rules for weighting the system

- Passing notes
- Strong and weak beats: (a) (b) (c)



- Chord notes
- Top or bottom notes of the contour



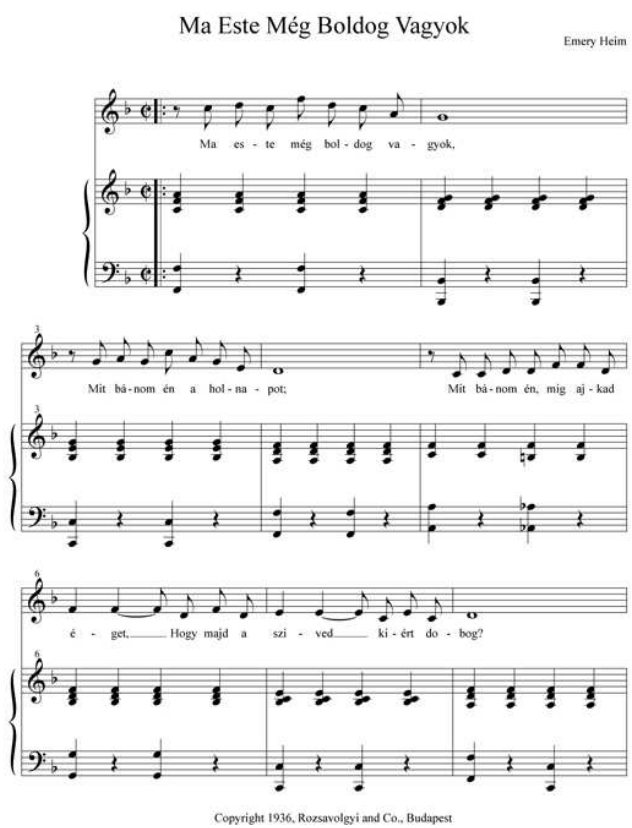
System for Near-Duplicate Identification

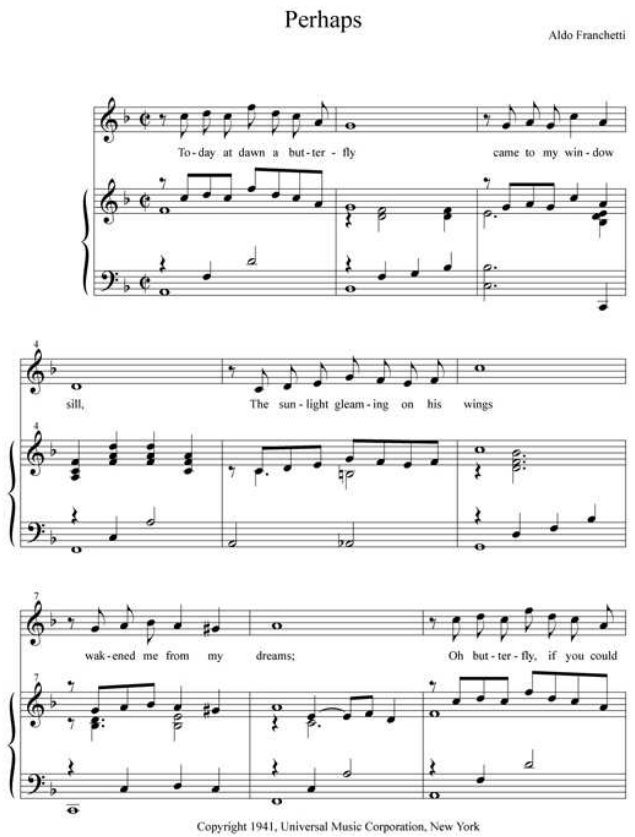
- Similarity of melodies
- System evaluated with excellent results in MIREX 2006 for monophonic similarity
- Columbia Law School: Music Plagiarism Project
- Subset of the RISM database as in MIREX

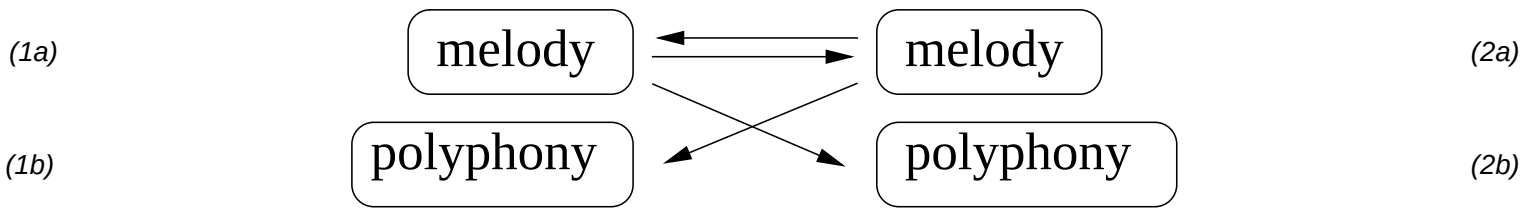
→ Can we retrieve an identified plagiarism in a such database ?

→ What is the score compared to non-plagiarism cases ?

An example: Heim v. Universal Pictures (1946)







Near-Duplicate Identification: Results

Query	**rank** 1 **score** 1	**rank** 2 **score** 2	**rank** 3 **score** 3
Heim vs Universal (1946)			
Vagyok	**Vagyok** 248.6	Perhaps 123.5	X 92.8
Perhaps	**Perhaps** 215.5	Vagyok 123.5	X 76.8
R. Mack vs G. Harrison (1976)			
Sweet Lord	**Sweet Lord** 178.9	So Fine 83.0	X 77.5
So Fine	**So Fine** 199.7	Sweet Lord 83.0	X 75.3
Selle vs Gibb (1984)			
Let It End	**Let It End** 192.4	How Deep 118.1	X 68.9
How Deep	**How Deep** 202.8	Let It End 118.1	X 83.8

Results for a few music copyright infringement cases with a database of 1650 incipits

Near-Duplicate Identification: Results

Query	**rank** 1 **score** 1	**rank** 2 **score** 2	**rank** 3 **score** 3
Heim vs Universal (1946)			
Vagyok	Vagyok 248.6	**Perhaps** 123.5	X 92.8
Perhaps	Perhaps 215.5	**Vagyok** 123.5	X 76.8
R. Mack vs G. Harrison (1976)			
Sweet Lord	Sweet Lord 178.9	**So Fine** 83.0	X 77.5
So Fine	So Fine 199.7	**Sweet Lord** 83.0	X 75.3
Selle vs Gibb (1984)			
Let It End	Let It End 192.4	**How Deep** 118.1	X 68.9
How Deep	How Deep 202.8	**Let It End** 118.1	X 83.8

Results for a few music copyright infringement cases with a database of 1650 incipits

Near-Duplicate Identification: Results

Query	**rank** 1 **score** 1	**rank** 2 **score** 2	**rank** 3 **score** 3
Heim vs Universal (1946)			
Vagyok	Vagyok 248.6	Perhaps 123.5	X 92.8
Perhaps	Perhaps 215.5	Vagyok 123.5	X 76.8
R. Mack vs G. Harrison (1976)			
Sweet Lord	Sweet Lord 178.9	So Fine 83.0	X 77.5
So Fine	So Fine 199.7	Sweet Lord 83.0	X 75.3
Selle vs Gibb (1984)			
Let It End	Let It End 192.4	How Deep 118.1	X 68.9
How Deep	How Deep 202.8	Let It End 118.1	X 83.8

Results for a few music copyright infringement cases with a database of 1650 incipits

Conclusion and Perspectives

- String matching algorithms improved for music documents
- Music similarity estimation useful for plagiarism detection

- Other music rules
- Polyphonic / polyphonic comparison improvement
- Multi-level approach
- Normalization of similarity scores

$\Longrightarrow$ Automatic plagiarism detection

Conclusion and Perspectives

- String matching algorithms improved for music documents
- Music similarity estimation useful for plagiarism detection

- Other music rules
- Polyphonic / polyphonic comparison improvement
- Multi-level approach
- Normalization of similarity scores

$\Longrightarrow$ Automatic plagiarism detection