Automated Music Transcription based on Formal Language Models

Florent Jacquemard
Lydia Rodriguez-de la Nava

Philippe Rigaux

le cnam

Masahiko Sakai

March 2023

Philippe Rigaux
le cnam Paris

Florent Jacquemard
Inria

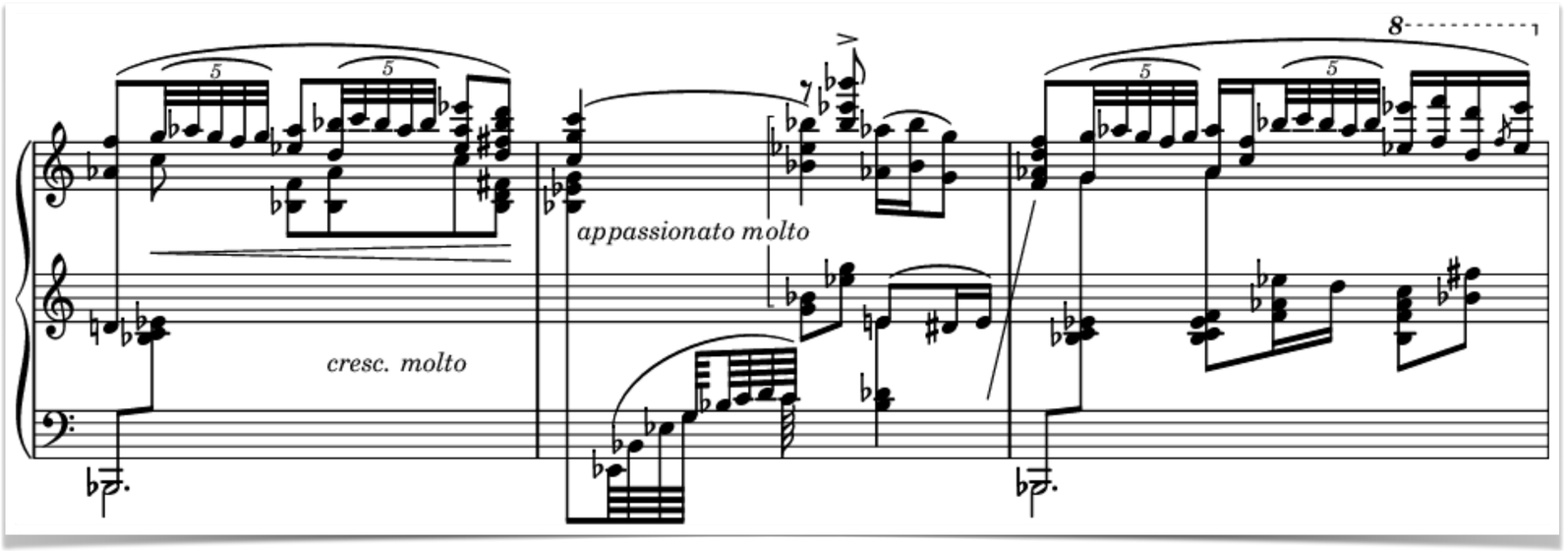
Raphaël Fournier-S'niehotta
le cnam

Lydia Rodriguez-de la Nava
PhD (Codex, Inria)

Tiange Zhu
PhD (Polifonia, H2020)

post-doc (Collabscore, ANR)

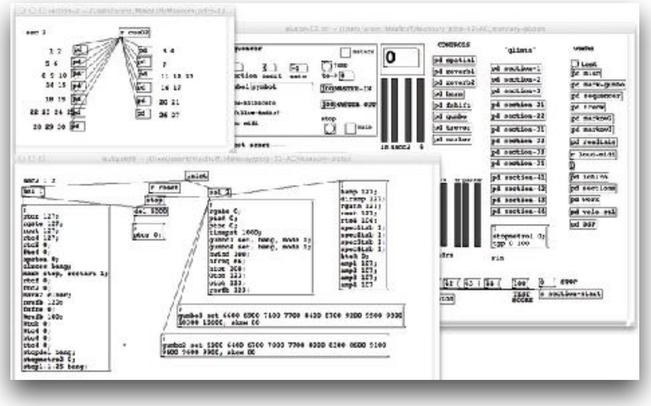
Music Notation Processing



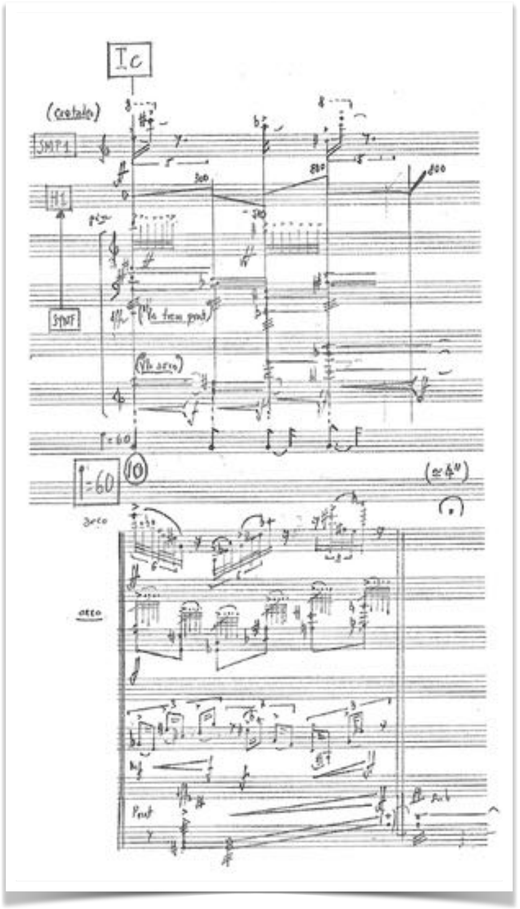
E. Granados, Goyescas
typesetted with Lilypond

Why studying Music Notation Processing?

Western Music Notation = graphical format for music practice,
in use since ~1000 years (Guido d'Arezzo)



or



Philippe Manoury
Tensio for string quartet and electronics

(digital) music scores, a natural language for

- performers
  performance : real-time reading or memoization
- composers
  authoring, exchange
- teachers & students
  transmission
- editors
  access digital score libraries *e.g.* nkoda.com
- librarians
  cultural heritage preservation: *e.g.* Gallica
- scholars (historians, musicologists…)
  research, analysis

Music Notation Processing by Vertigo & Codex

Structured Music Score models
hierarchical representation of music scores
finite representations of languages (*style*)

Search and Retrieval
indexing
exact and approximate
search and query

Similarity metrics
string and tree
edit-distances

H2020 Polifonia
U. Bologna
Open University
King's College London
Vrije U. Amsterdam

IReMus (Paris)
CNRS, Paris

ANR Collabscore
IRISA
French National Library
fondation Royaumont

JSPS 採譜
Nagoya U. (Sakai lab.)
JAIST (Tojo lab.),

AlgoMus
Lille, France

Digital Music Score Collections
Cultural heritage preservation and study
neuma.huma-num.fr

Computational Musicology

Optical Music Recognition
Crowdsourced correction

Automated Music Transcription
qparse

Tree-Structured Representation of Written Music

646 Nature Vol. 263 October 21 1976

articles

Perception of melodies

H. C. Longuet-Higgins

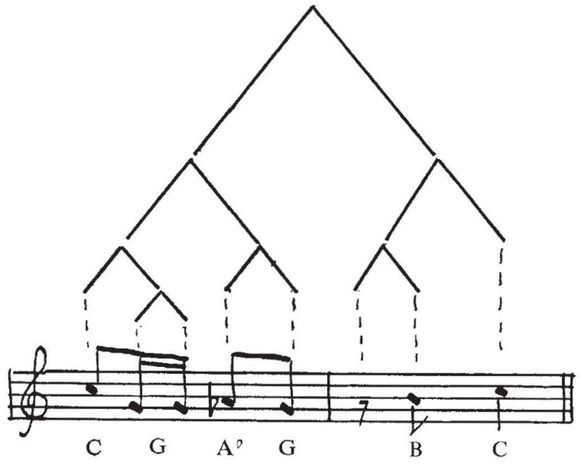
Centre for Research on Perception and Cognition, Laboratory of Experimental Psychology, University of Sussex, Brighton BN1 9QG, UK

A computer program has been written which will transcribe a live performance of a classical melody into the equivalent of standard musical notation. It is intended to embody, in computational form, a psychological theory of how Western musicians perceive the rhythmic and tonal relationships between the notes of such melodies.

A SEARCHING test of practical musicianship is the 'aural test' in which the subject is required to write down, in standard, musical notation, a melody which he has never heard before. His transcription is not to be construed as a detailed record of the actual performance, which will inevitably be more or less out of time and out of tune, but as an indication of the rhythmic and tonal relations between the individual notes. How the musical listener perceives these relationships is a matter of some interest to the cognitive psychologist. In this paper I outline a theory of the perception of classical Western melodies, and describe a computer program, based on the theory, which displays, as best it can, the rhythmic and tonal relationships between the notes of a melody as played by a human performer on an organ console.

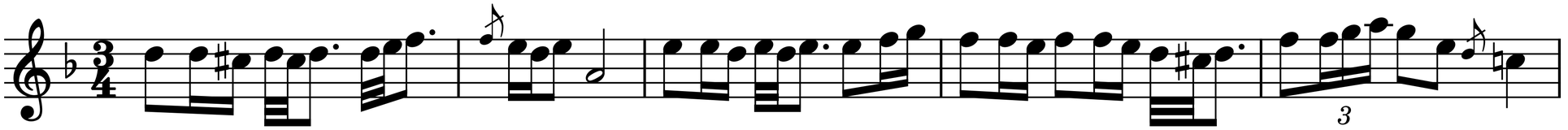
The basic premise of the theory is that in perceiving a melody the listener builds a conceptual structure representing the rhythmic groupings of the notes and the musical intervals between them. It is this structure which he commits to memory, and which subsequently enables him to recognise the tune, and to reproduce it in sound or in writing if he happens to be a skilled musician. A second premise is that much can be learned about the structural relationships in any ordinary piece of music from a study of its orthographic representation. Take, for example, the musical cliché notated in Fig. 1.

Fig. 1



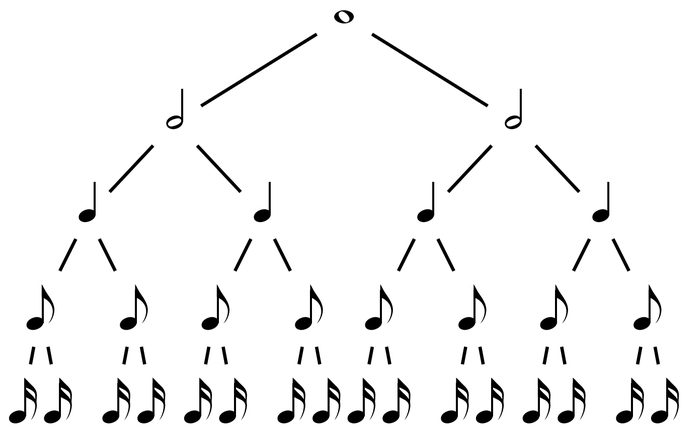
Polonaise in D minor from Notebook for Anna Magdalena Bach BWV Anh II 128

beamed



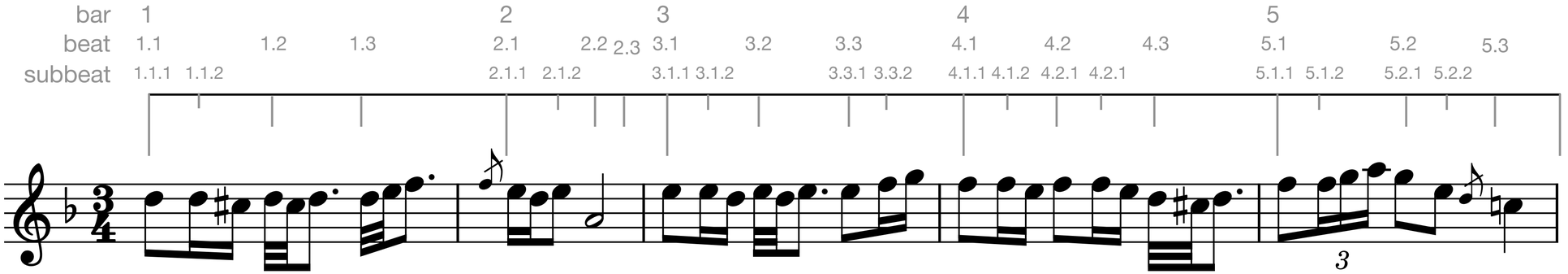
unbeamed

hierarchical note durations



Polonaise in D minor from Notebook for Anna Magdalena Bach BWV Anh II 128

metric structure



beamed

unbeamed

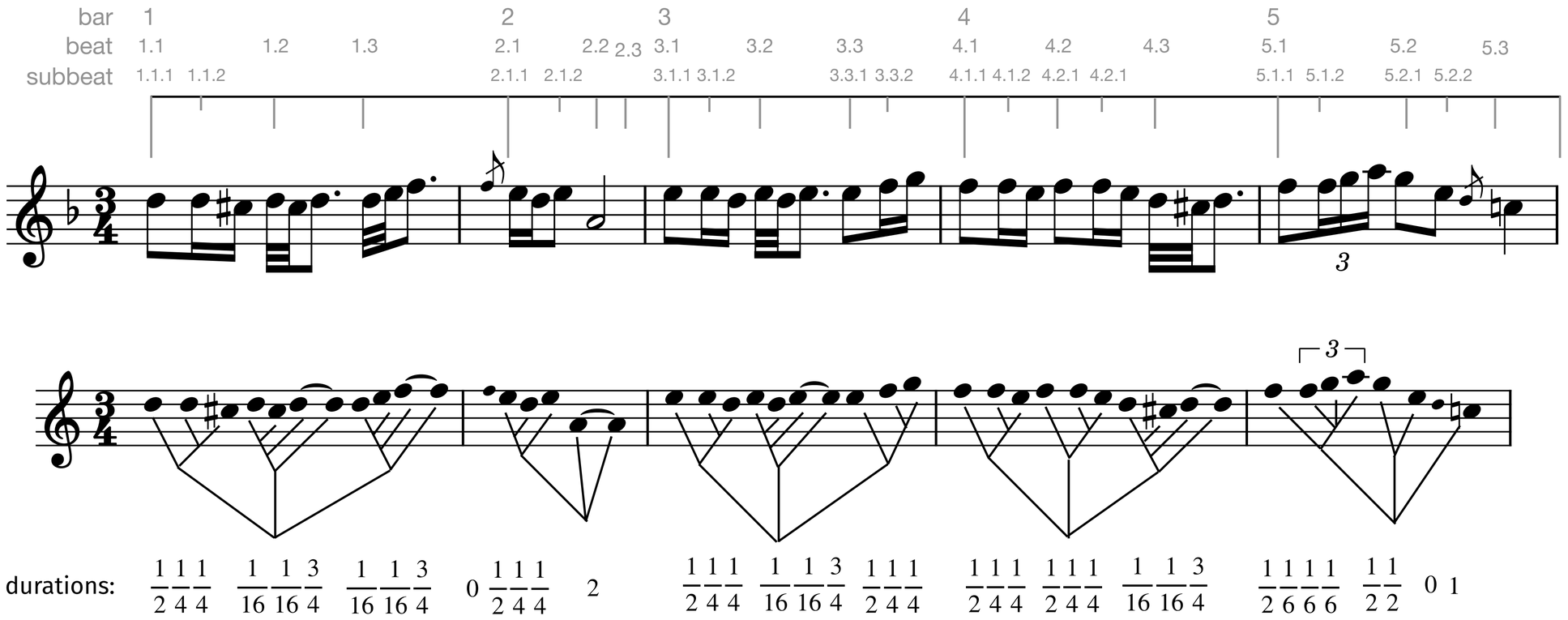
grouping notes with measure bars and beams

- eases readability (player reads in a real-time context)

- highlight the metric structure
  hierarchy of strong / weak beats

Polonaise in D minor from Notebook for Anna Magdalena Bach BWV Anh II 128

metric structure

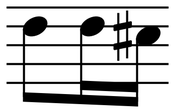


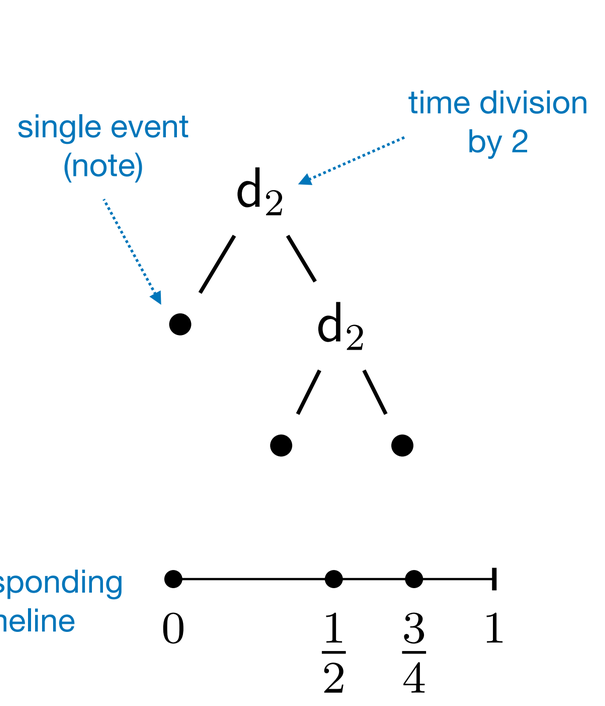
Tree-structured Representation of Music Notation : Rhythms

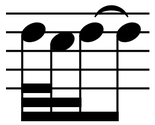
Labeled-Tree representation of the proportional rhythmic notation

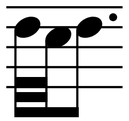
Hierarchical encoding of durations: “*the* (duration) *data is in the structure*” :

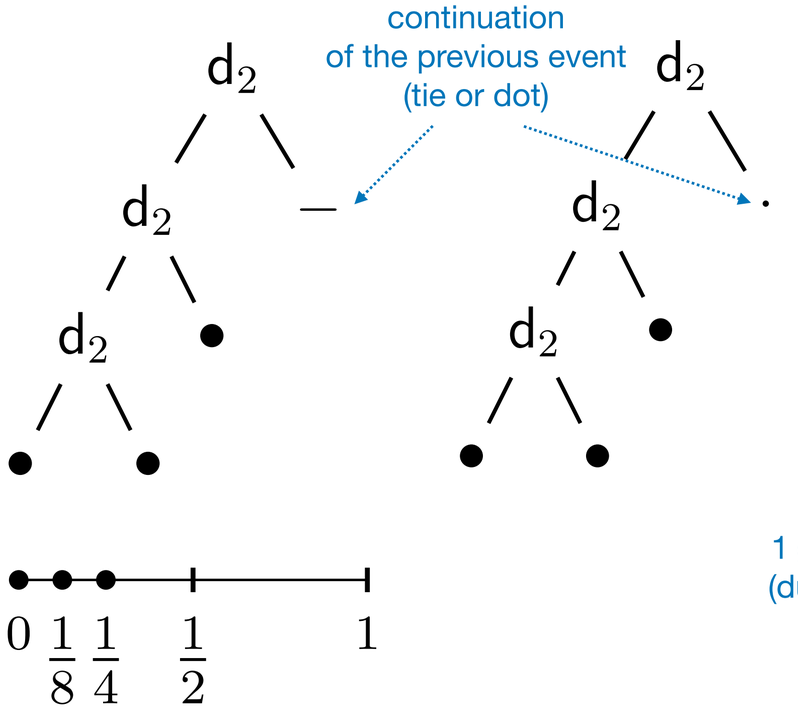
- the tree leaves contain the events
- the branching define durations, by partitioning of time intervals

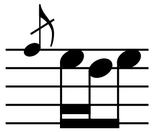


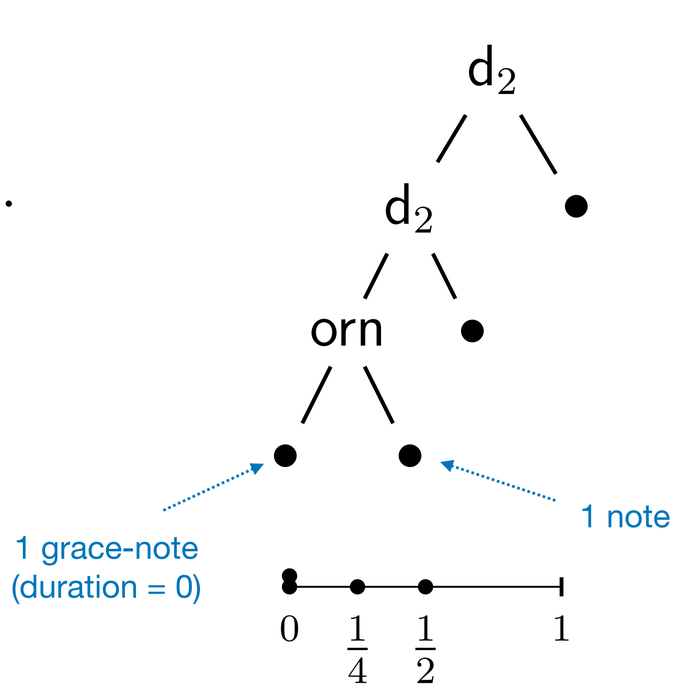




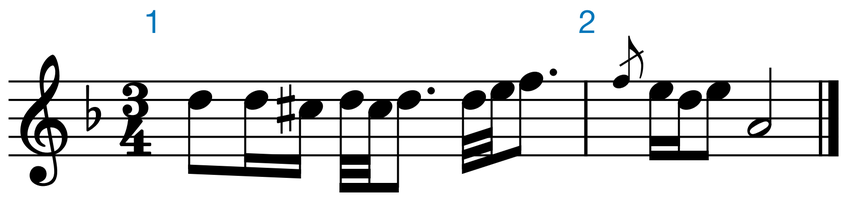


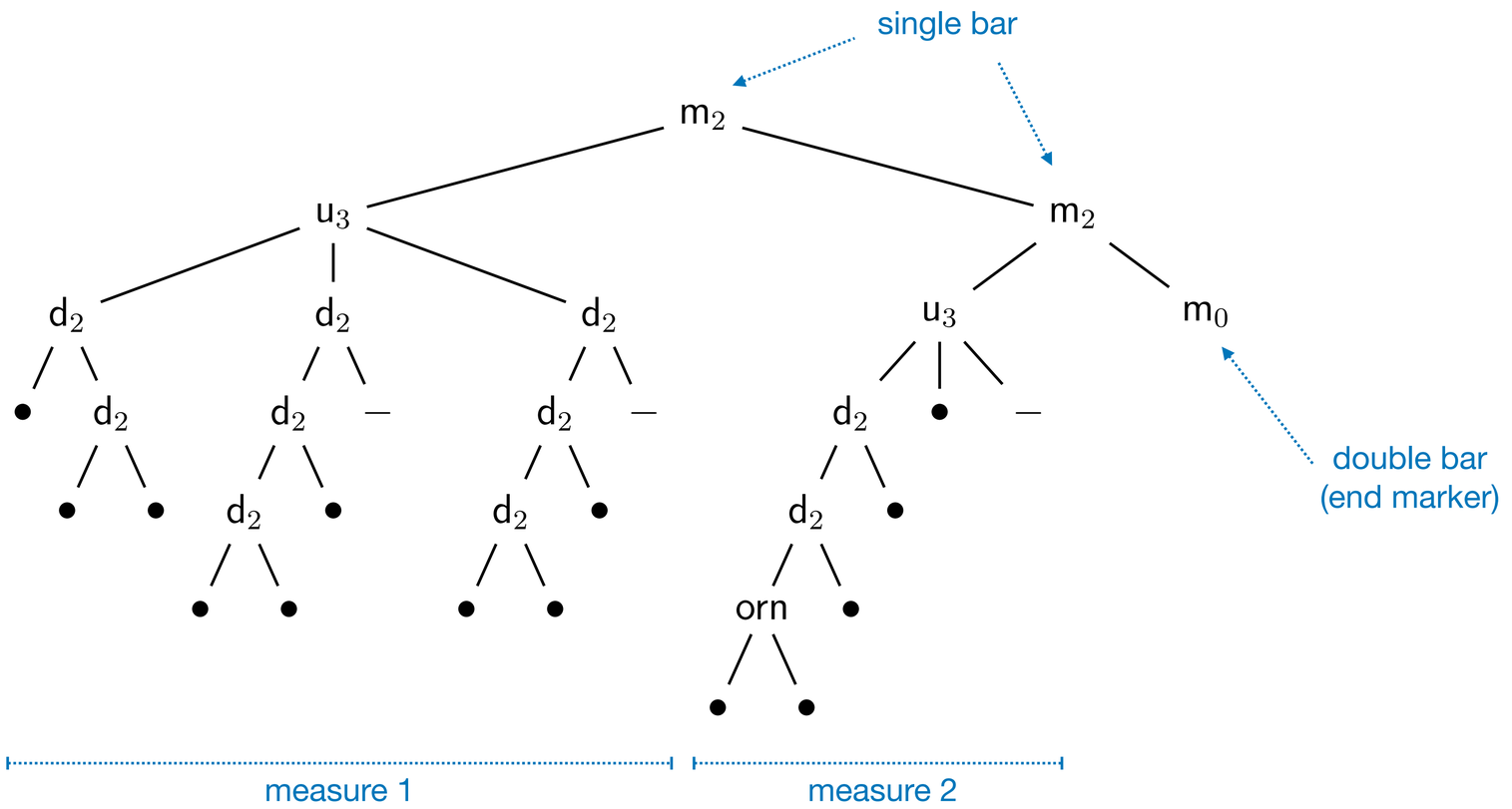






Tree-structured Representation of Music Notation : Measures

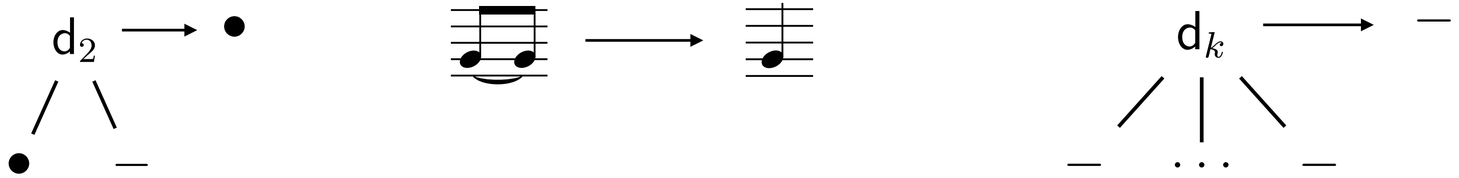




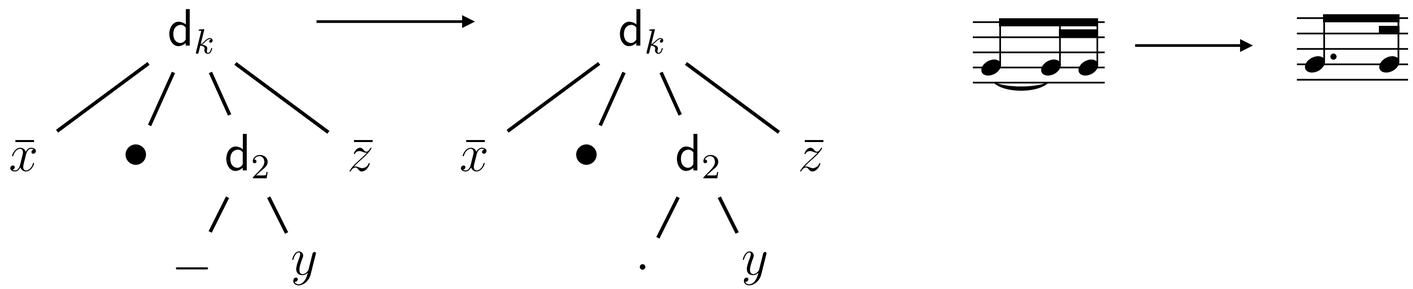
Term Rewriting Rules

for the transformation of score representations

tie simplification



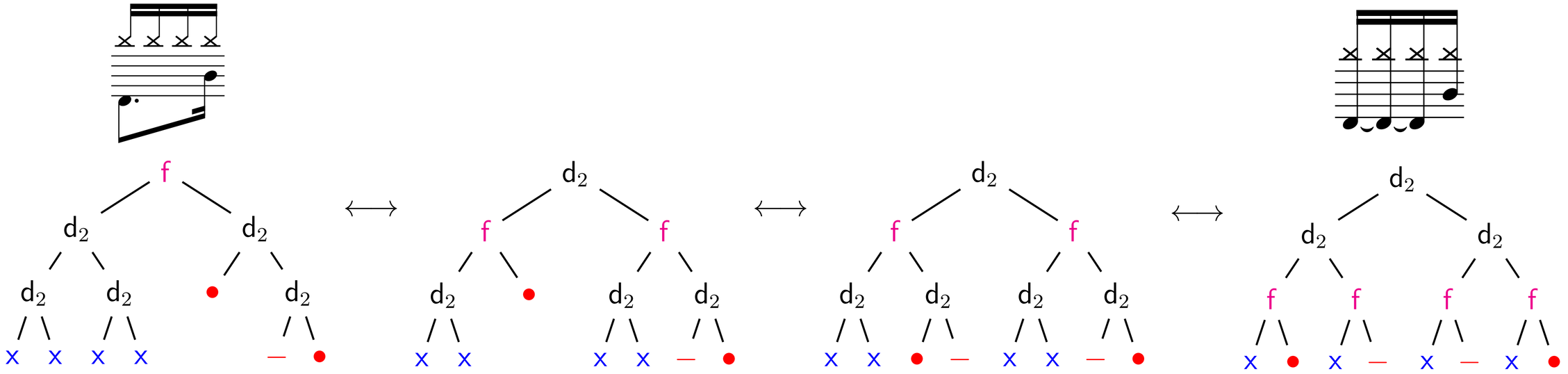
tie to dot



staccato

Tree-structured Representation of Music Notation : Voices

form symbol f
term rewriting: rules swapping f and tuple symbols



Automatic Music Transcription based on Weighted Parsing

Automated Music Transcription today

Conversion of a recorded music performance into a music score ~ *speech-to-text* in NLP

source(s)

Audio recording

MIDI device
(score edition)

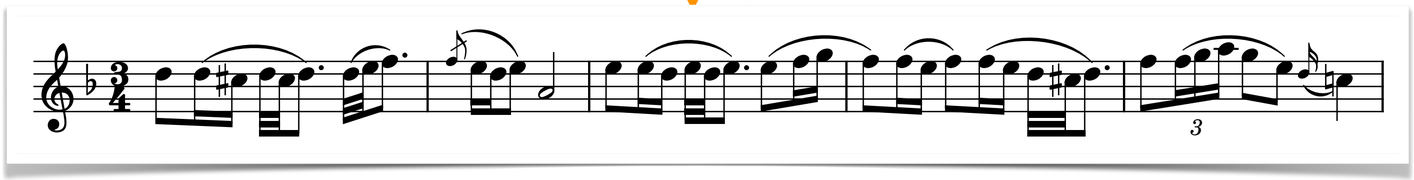
Algorithmic composition
DAW

audio Music Information Retrieval
- fundamental freq. estimation
- onset detection
- beat tracking …

intermediate representation
piano roll (MIDI file)
- reals-time durations (*seconds*), unquantized
- quantized pitches

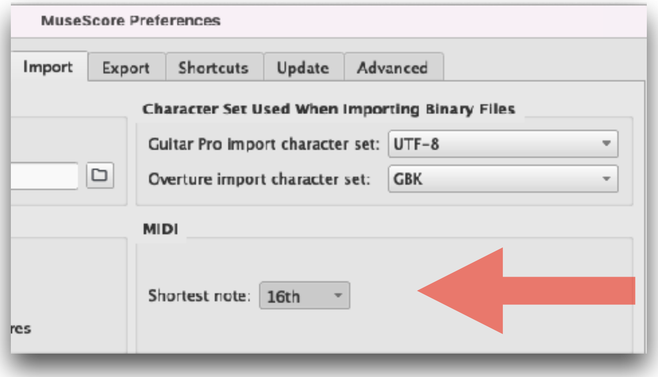
symbolic Music Information Retrieval
- rhythm quantization
- tempo tracking
- score engraving…

target music score
(e.g. XML file)
- musical time durations (*beats*) quantized

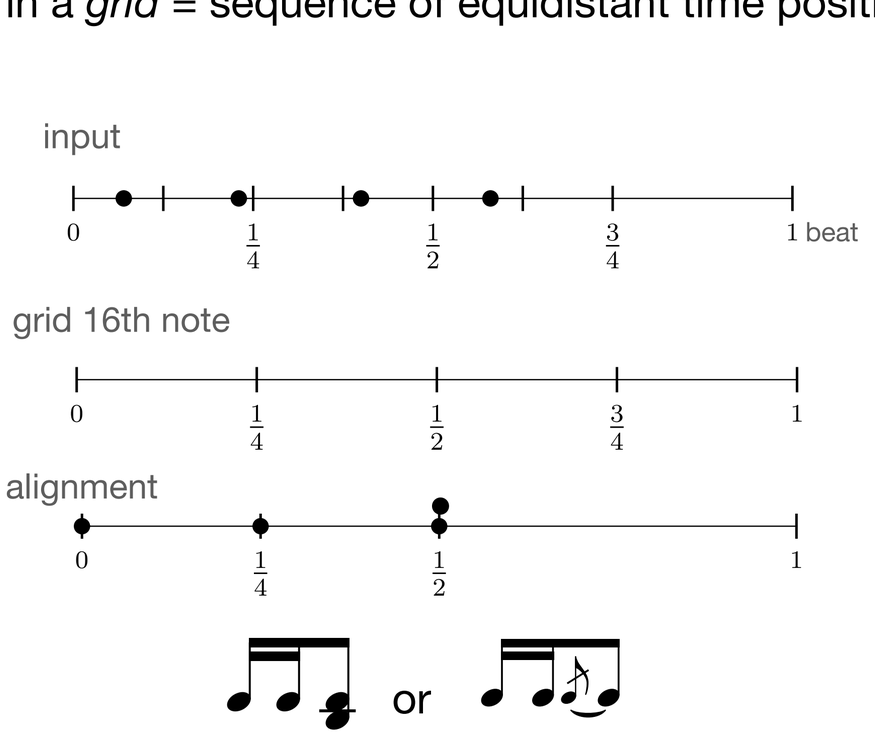


Grid based Approaches to Rhythm Quantization

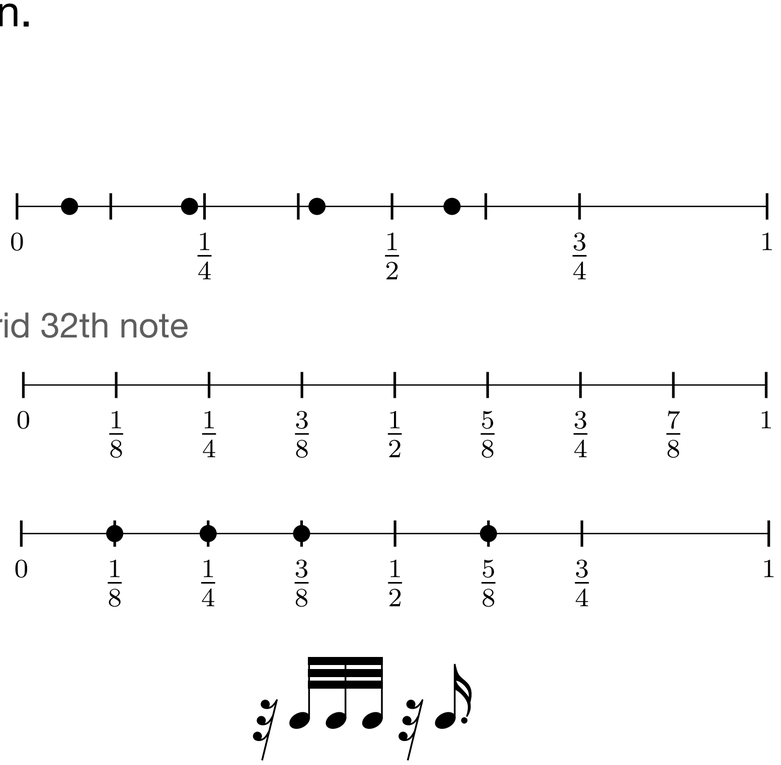
Rhythm quantization with grids, *e.g.* MIDI files import
- in score editors (Finale, Sibelius, Dorico, Musescore…),
- or in DAWs (Ableton Live, Logic…)



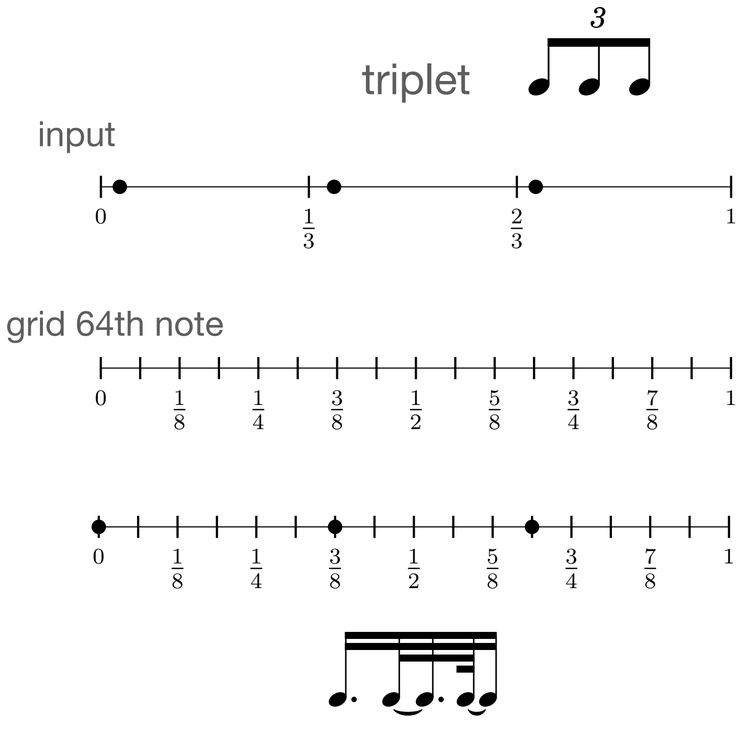
Alignment of every input time point (onset) to the closest position in a *grid* = sequence of equidistant time position.



poor fit, good readability

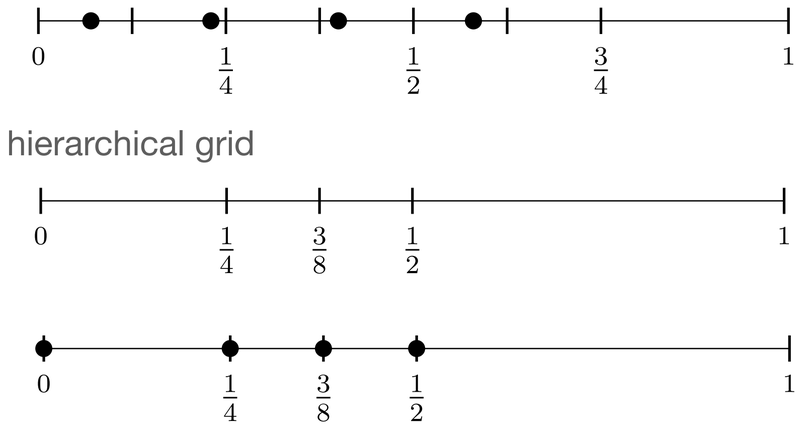


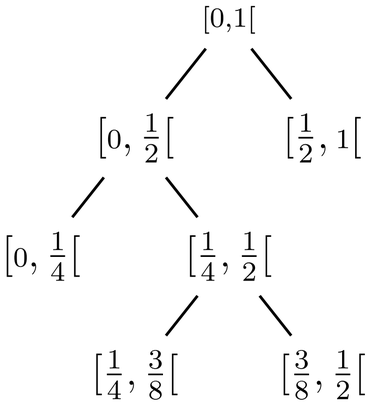
good fit, bad readability

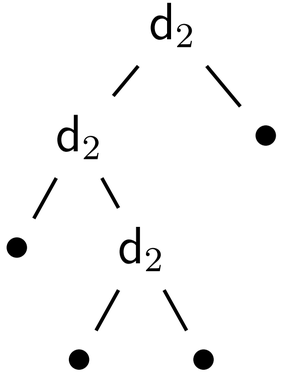


good fit, bad readability

Hierarchical (irregular) Grids







closer to intuition

regular grids

- search of a best quantization is possible by a brute-force enumeration:
  8th note grid, 16th, 32th, 64th…
- result not always optimal
- problems with tuplets (so called “*irrationals”* 3, 5, 7…*)*

hierarchical grids

- more “natural” results
- brute force enumeration impossible
- how to specify the grids to try ?

Transcription as Parsing

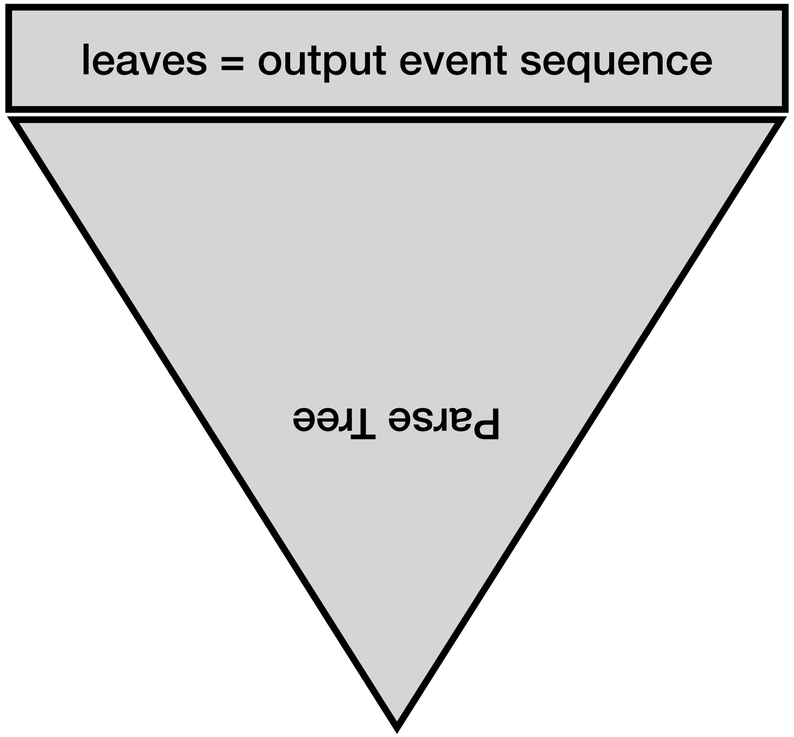
piano roll
= sequence of timestamped input events

parsing

structuring a linear representation according to a language model

tree-structured representation
of an output **music score**

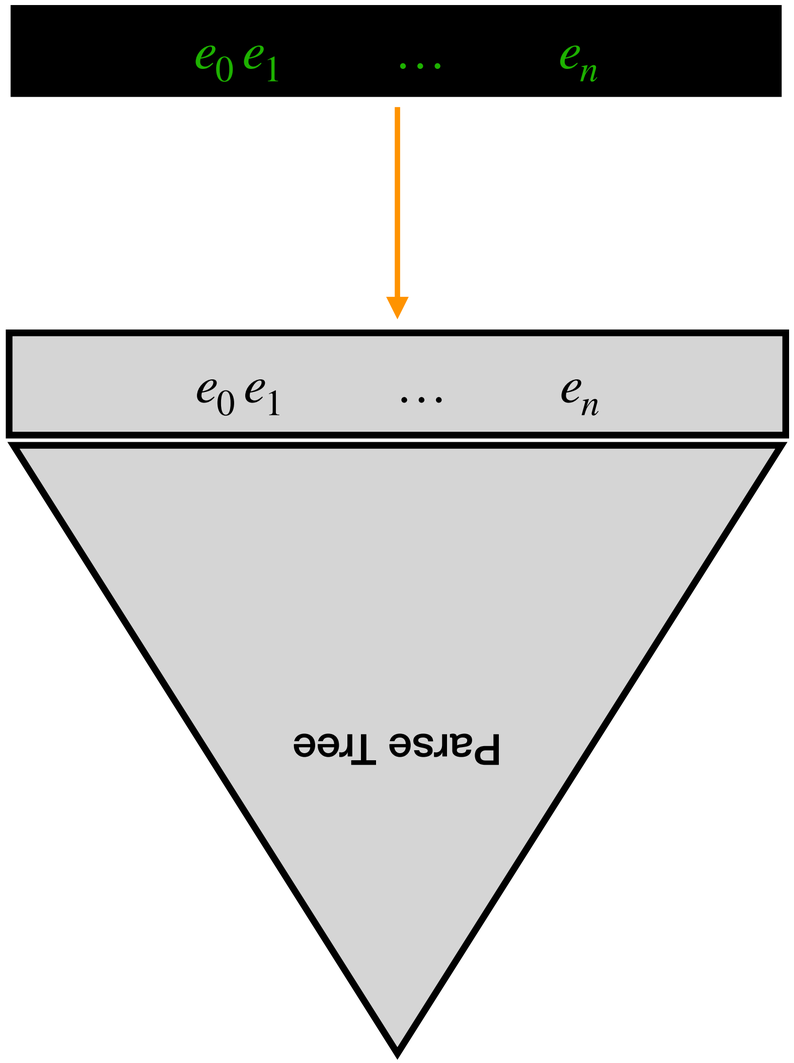
conforming to a
prior language (expected notation)
= **Tree Grammar** $\mathscr{G}$



extensions of parsing are needed
for the case music transcription:
- weighted extension
- symbolic weighted extension
  (*quantitative parsing*)

Objective of **parsing** : finding a parse tree of $\mathscr{G}$ that yields $e_0\, e_1 \ldots e_n$

input sequence
(*events*)



yield sequence
(leaves)

With an ambiguous prior grammar $\mathscr{G}$ there might exist several parse trees (exponentially many).

In order to choose one (or some) parse trees, they are **ranked** according to **weight** values, computed by **Weighted Tree Grammar**

In general, the weight values are taken in a **commutative Semiring** $\langle \mathbb{S}, \oplus, \mathbb{0}, \otimes, \mathbb{1} \rangle$

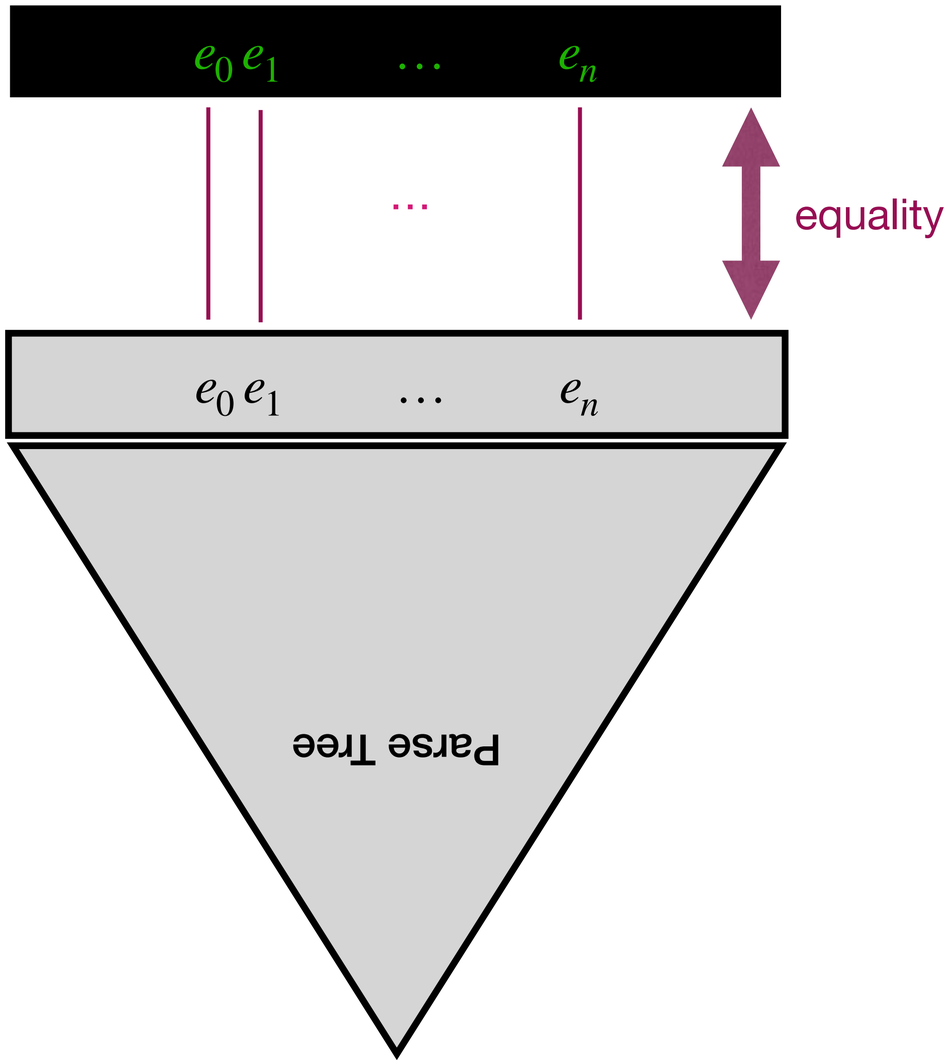
	domain	$\oplus$	$\otimes$	$\mathbb{0}$	$\mathbb{1}$
Boolean	$\{ \bot , \top \}$	$\vee$	$\wedge$	$\bot$	$\top$
Viterbi	$[0,1] \subset \mathbb{R}$	max	$\times$	0	1
Tropical min-plus	$\mathbb{R}_+ \cup \{+\infty\}$	min	$+$	$+\infty$	0

Objective of **weighted parsing** : find the best (*wrt* weight) parse tree of $\mathscr{G}$ that yields $e_0\, e_1 \ldots e_n$

(Weighted) Parsing : I-O relationship

Objective of **weighted parsing** : find the best (*wrt* weight) parse tree of $\mathcal{G}$ that yields $e_0\, e_1 \ldots e_n$

input sequence
(*events*)



yield sequence
(*tokens*)

trivial (1-1) correspondence between
input string and
leaves of output tree

Tokenisation

a more involved tokenisation process,
several correspondences possible

ko ko de | Ha ki mo no wo | nu i de | ku da sa ï

here shoes remove please

ここではきものをぬいでください

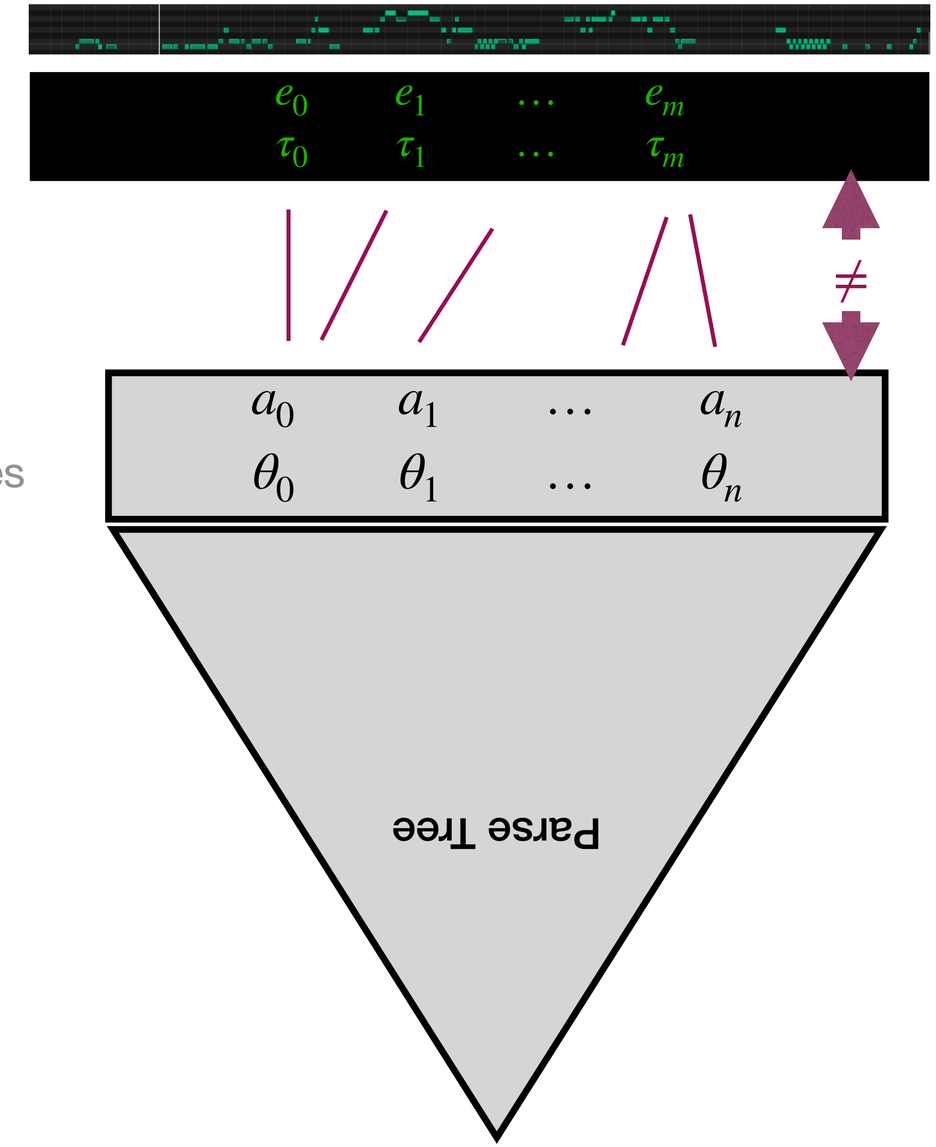
here clothes remove please

ko ko de wa | ki mo no wo | nu i de | ku da sa ï

in music, the correspondance between the (MIDI) input sequence and sequence of output leaves is also not 1-1

Quantitative Parsing (extension 2) : IO measure

input sequence



yield
= sequence of leaves

we extend weighted parsing by
ranking solutions with:

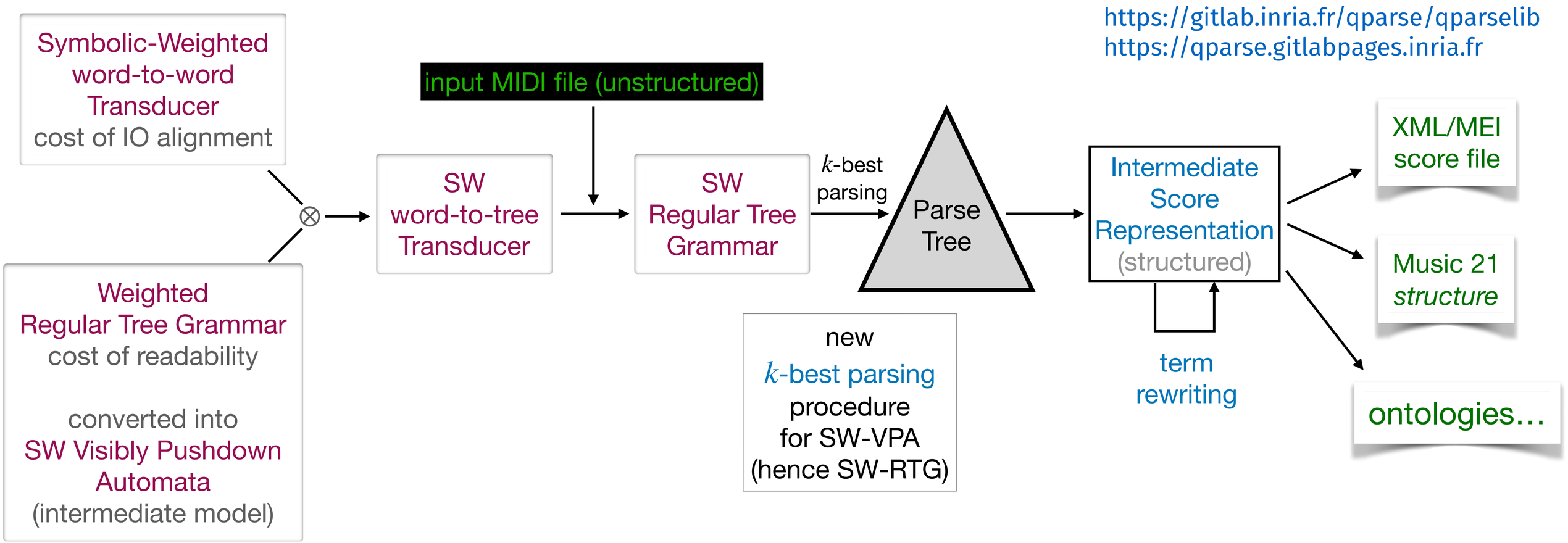
a measure of input / output fitness
= cost of IO alignement
computed by a **Weighted Transducer**

⊗

measure of cost-to-read
weight value
computed by the **Weighted Tree Grammar**

Objective of **quantitative parsing** : find a parse tree of $\mathscr{G}$ minimizing the product of I/O alignment cost and weight.

Automated Music Transcription: qparse framework



qparse (75 Kloc C++) is an implementation of

- the above transcription by parsing framework
- an intermediate score model
- other subtasks: pitch-spelling, key estimation, beat tracking…

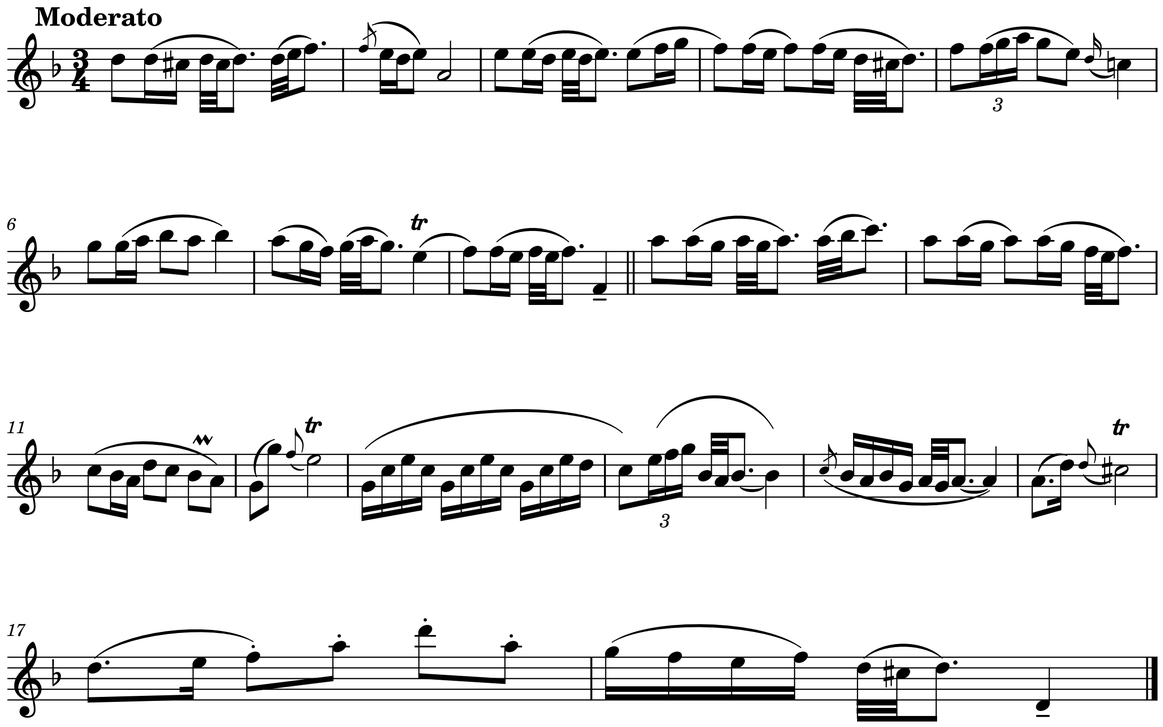
- command lines tools: `monoparse`, `drumparse`, grammar-learning, engraving (from quantified input)
- Python binding - Lydia Rodrigez-de la Nava evaluation scripts
- online port, real-time - Leyla Villaroel

Results, Datasets

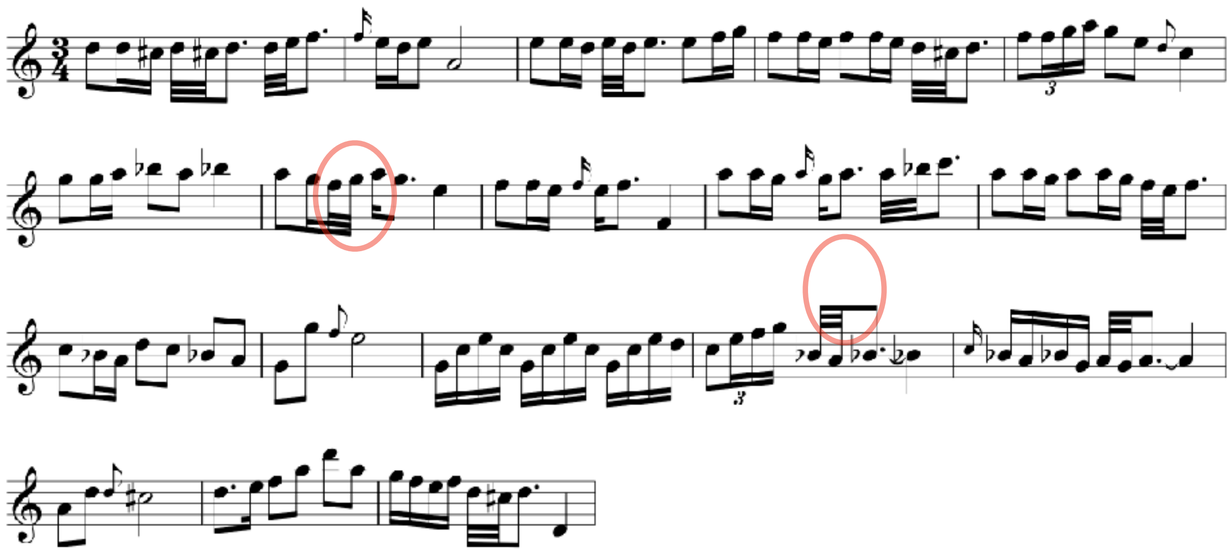
Monophonic transcription

monophonic : one note at a time
Good results for complex cases (ornaments, mixed tuplets, mixed note durations, silences…)
~ 100ms for the transcription of 1 score

original score



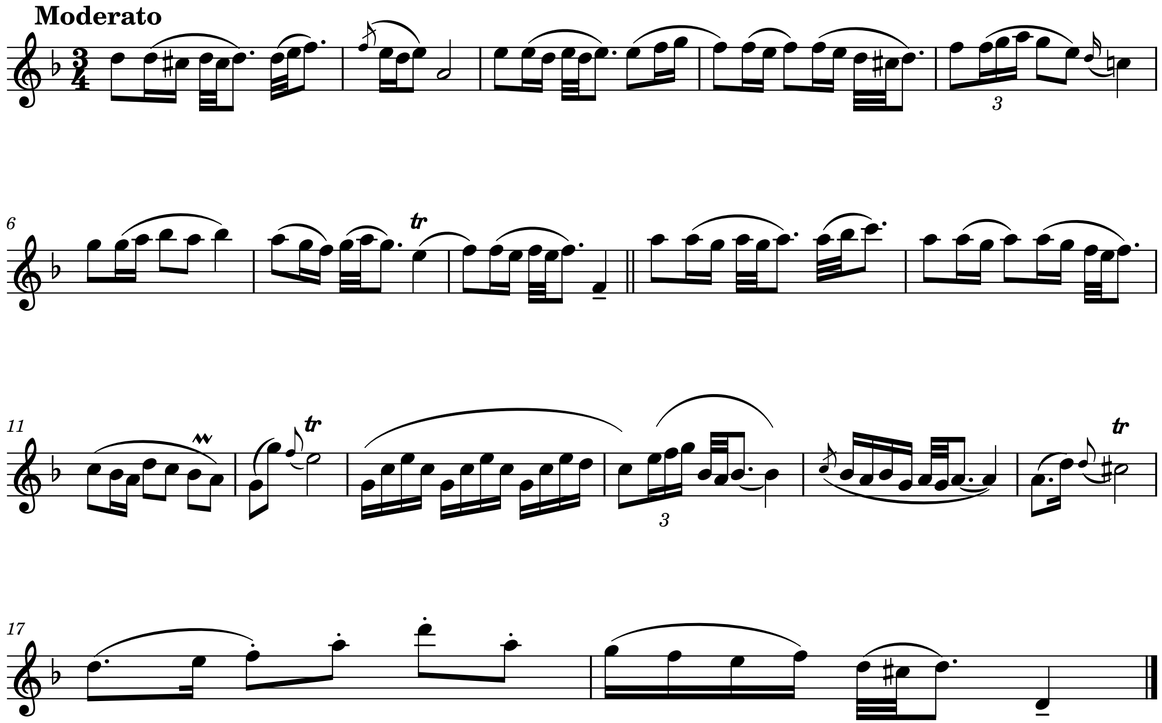
transcription of MIDI recording by qparse



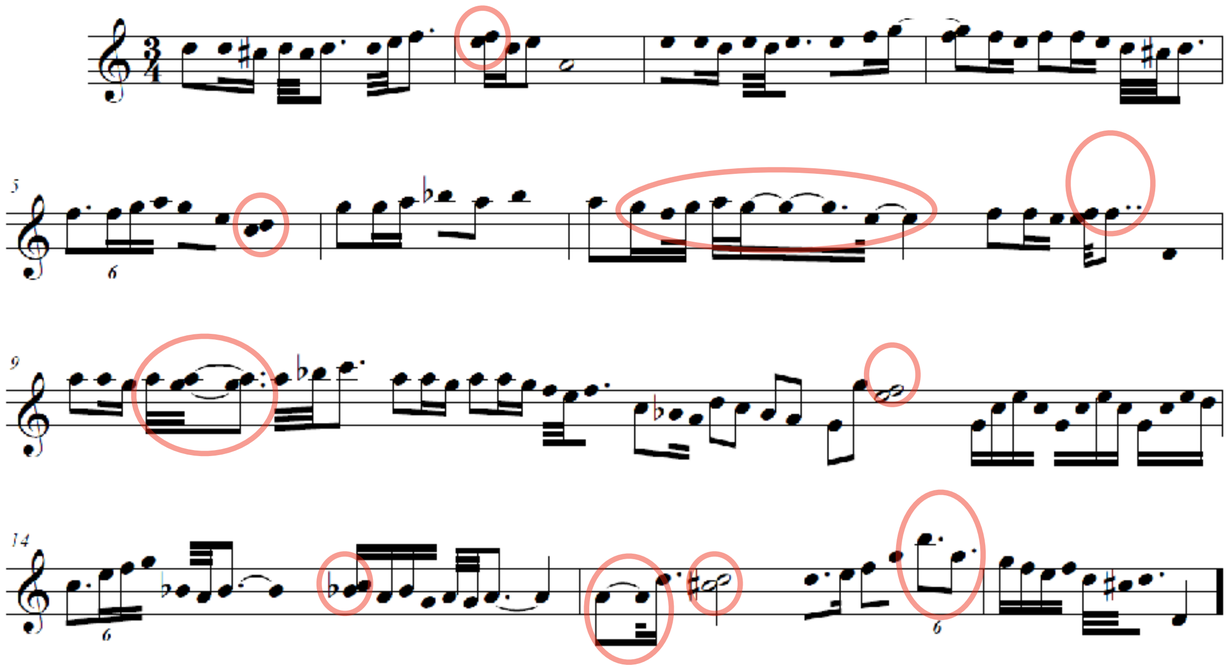
Polonaise in D minor from Notebook for Anna Magdalena Bach BWV Anh II 128

Monophonic transcription

original score



transcription of MIDI recording by Finale



Polonaise in D minor from Notebook for Anna Magdalena Bach BWV Anh II 128

Evaluation data for Monophonic transcription

Evaluation and calibration of transcription tools

Lamarque-Goudard dataset
w. Francesco Foscarin, Teysir Baoueb

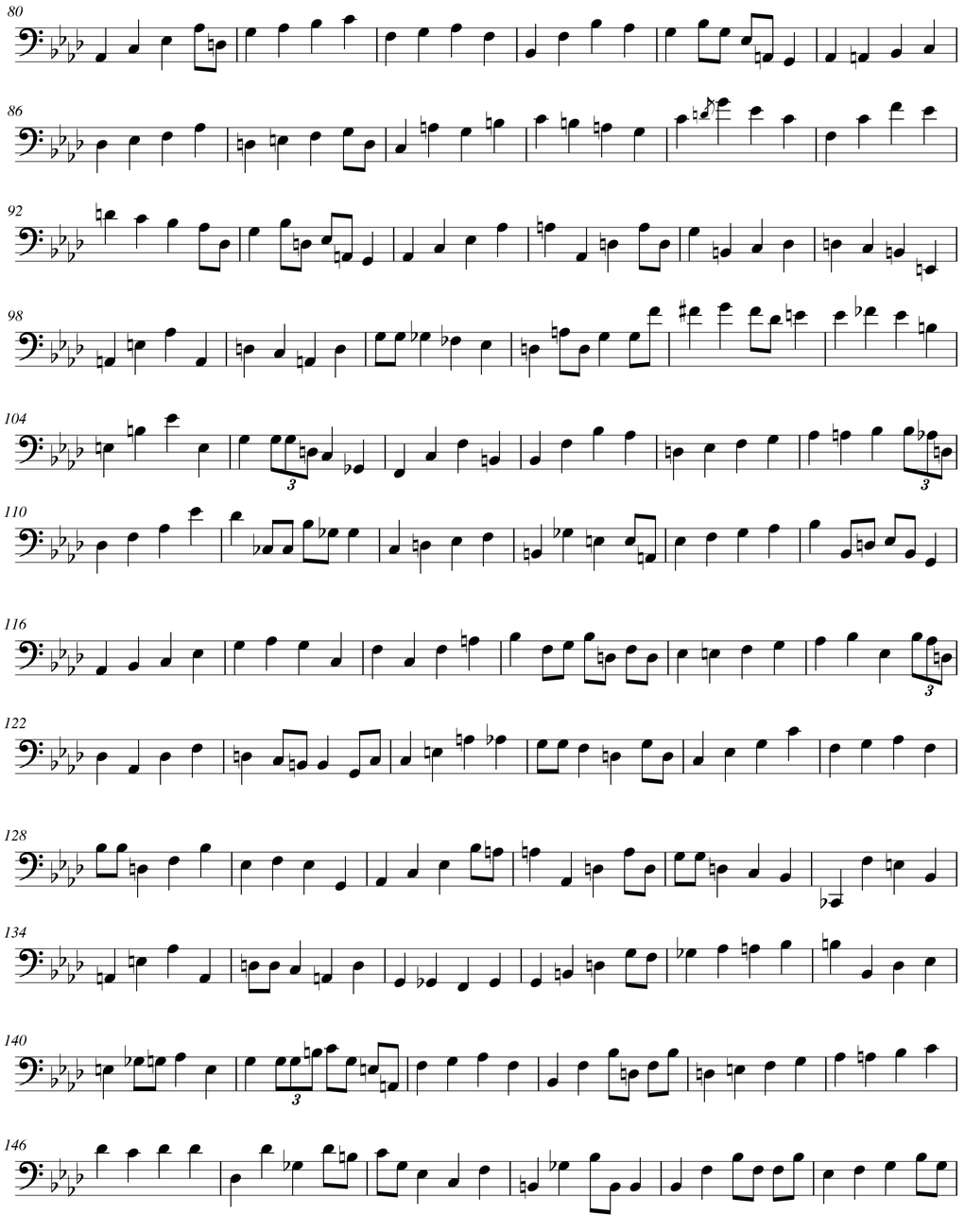
- 283 monophonic extracts
  of classical repertoire
  inspired by a rhythm learning method
- ~ 20 measures per extract
- progressive difficulty
  cover a very large spectrum of rhythmic features
- score files (XML) and MIDI performances

Generation of artificial performances
Madoka Goto, Masahiko Sakai (Nagoya U.), Satoshi Tojo (JAIST)

- analyses from the GTTM database (time-span trees)
- score segmentation (phrasing), according to time-span trees
- performance generation by Director Musices (Anders Friberg)

FiloBass by John-Xavier Riley (C4DM, QMUL)
project “*Dig That Lick*”

- jazz bass lines, acc. of saxophone
- 48 tracks,
  24 recorded hours of melodies and improvisations
- qparse as backend of an audio-to-MIDI
  transcription procedure
- prior beat (measure) tracking



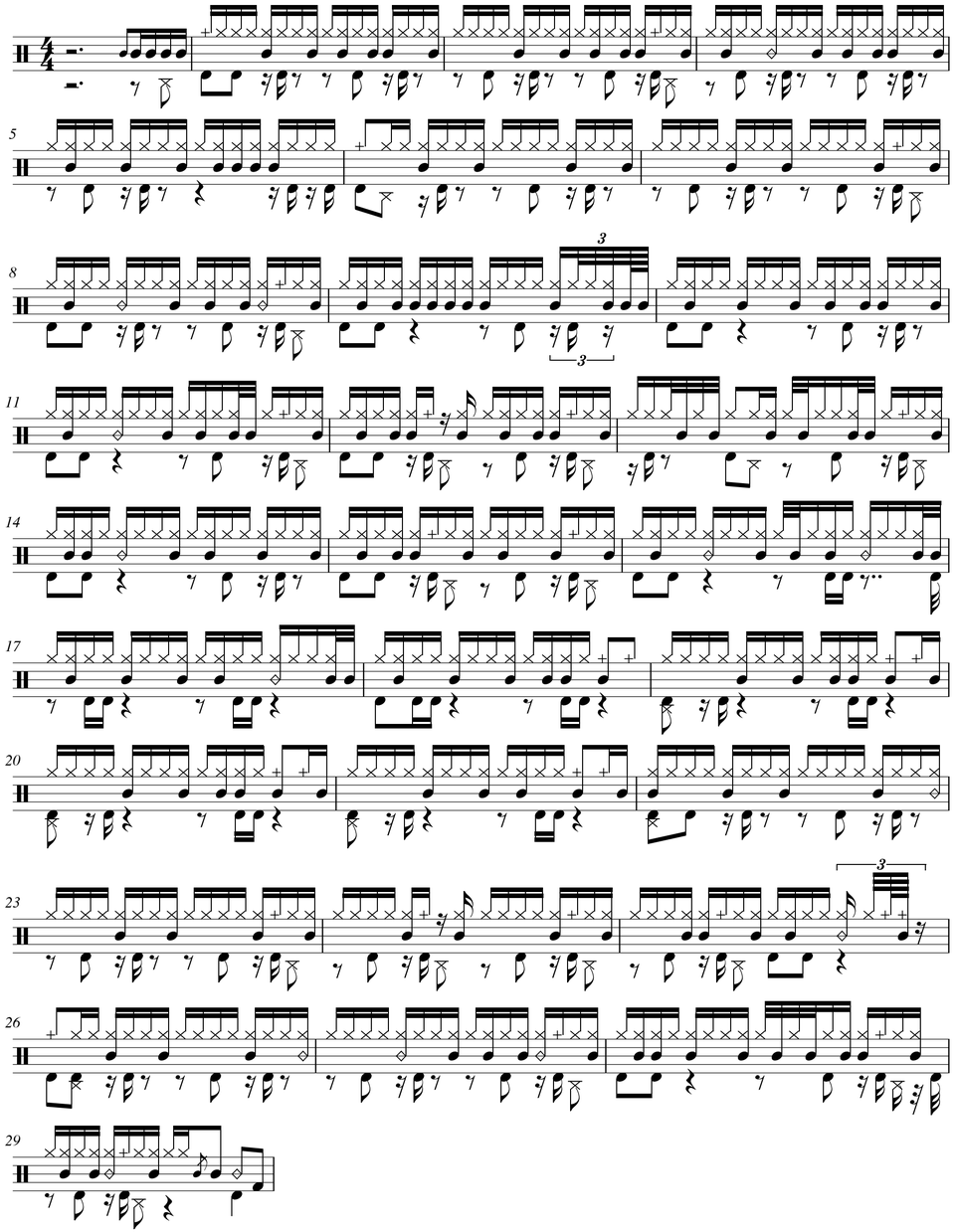
Groove MIDI Dataset

- by Google Magenta
- 13.6 hours, 1150 MIDI files, ~ 22000 measures recorded by professional drummers on a electronic drum kit
- audio (wav) files synthesized from (and aligned to) MIDI files for evaluation of audio-to-MIDI drum transcription
- no score files!

Scoring the GMD with qparse

Martin Digard (INALCO)

- all score files (XML) produced from the MIDI files with the same generic tree grammar (4/4 measure)
- polyphonic case-study, simpler than piano
- specific drumming constraints (hands ≤ 2, feet ≤ 2)
- processing errors from MIDI sensors



Piano & Guitar transcription

From Monophonic to Polyphonic Transcription, stepwise:

- From Monophonic to **Homophonic Transcription** (chords)
  Yusuke (Nagoya U.)

- **Drum Transcription** Martin Digard, Lydia Rodrigez-de la Nava
  Google GMD

- **Voice separation**
  Lydia Rodrigez-de la Nava, Augustin Bouquillard
  integration for piano guitar transcription:
  - before parsing, or
  - after parsing (on intermediate model), or
  - joint with parsing.

- **Dataset ASAP** - Francesco Foscarin, Andrew Mc Leod
  MIDI and audio recording from Yamaha piano competition
  + XML scores
  + alignments
  + beat tracking annotations

Conclusion

MIDI-to-Score Automated Music Transcription approach

- based on quantitative parsing techniques
  and Symbolic Weighted formal language formalisms
  (*Tree Automata* and *word-to-word Transducers*)
- with prior language of notation *style*
  and prior IO measure
- (abstract) hierarchical score model
  as intermediate representation for score generation

- can handle complex notation cases:
  ornaments, mixed tuplets, mixed note durations, silences…
- efficient

- case studies: Monophonic, Drums
- ongoing work on Polyphonic case studies: guitar, piano

ありがとうございます