Analysis of Music Data

Florian Schoppmann

Department of Computer Science

International Graduate School *Dynamic Intelligent Systems*
University of Paderborn

July 17, 2007

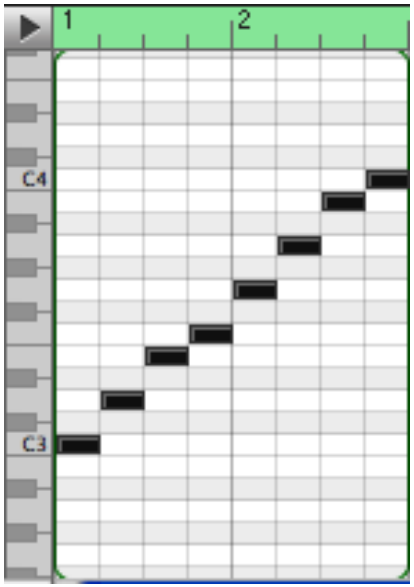
Analysis of Music Data

Goals:

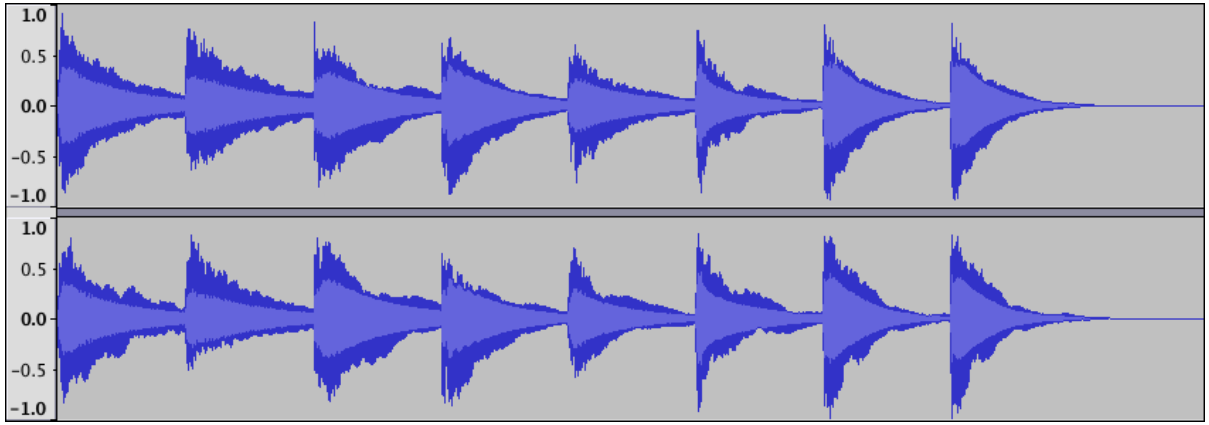
- Measuring Similarity of Music Data
  - Memory research
  - Music psychology
  - Music analysis ("ethnomusicology")
  - Copyright issues
- Extracting high-level information from general audio signals
  - Restoration of musical sources
  - Music transcription
- Sound characteristics of orchestra instruments
  - classicifaction of the register of an instrument by timbre (and not pitch)

Music Representation

Symbolically,
e.g., also MIDI:



Audio Signals, e.g., in AIFF, MP3, etc.:



MIDI Parameters: Note, Velocity, Modulation, Balance, etc.

Technical Terms

Pitch Perceived fundamental frequency of a sound

Fundamental Frequency f_0 lowest frequency in a harmonic series

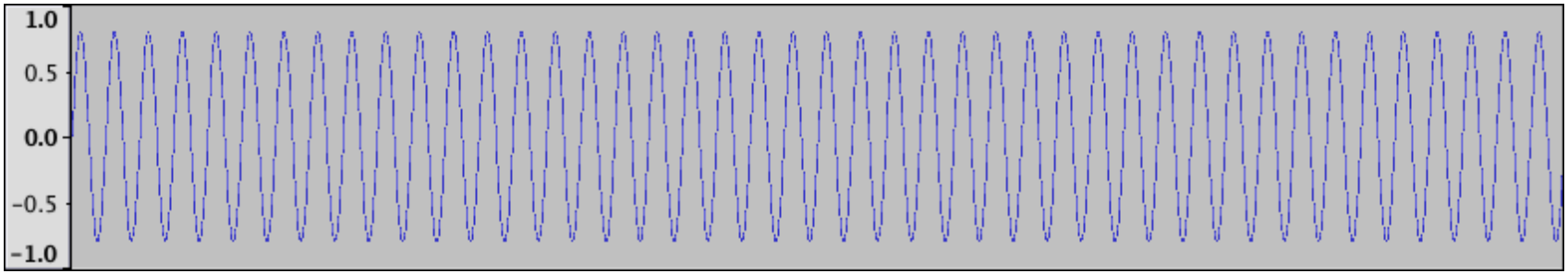
Harmonic Series (Besides maths...) Multiples of f_0

Overtone/Partial sinusoidal component of a waveform, of greater frequency than f_0

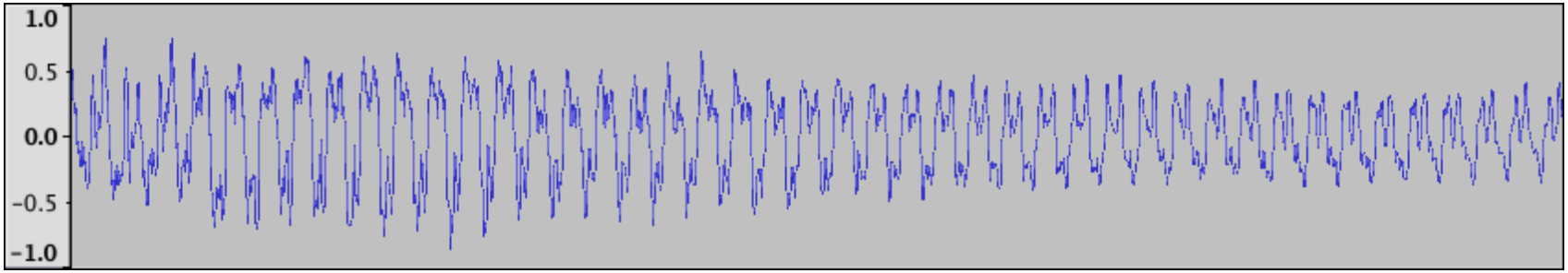
Register Relative "height" or range of a note, set of pitches, melody, instrument, etc.

Example

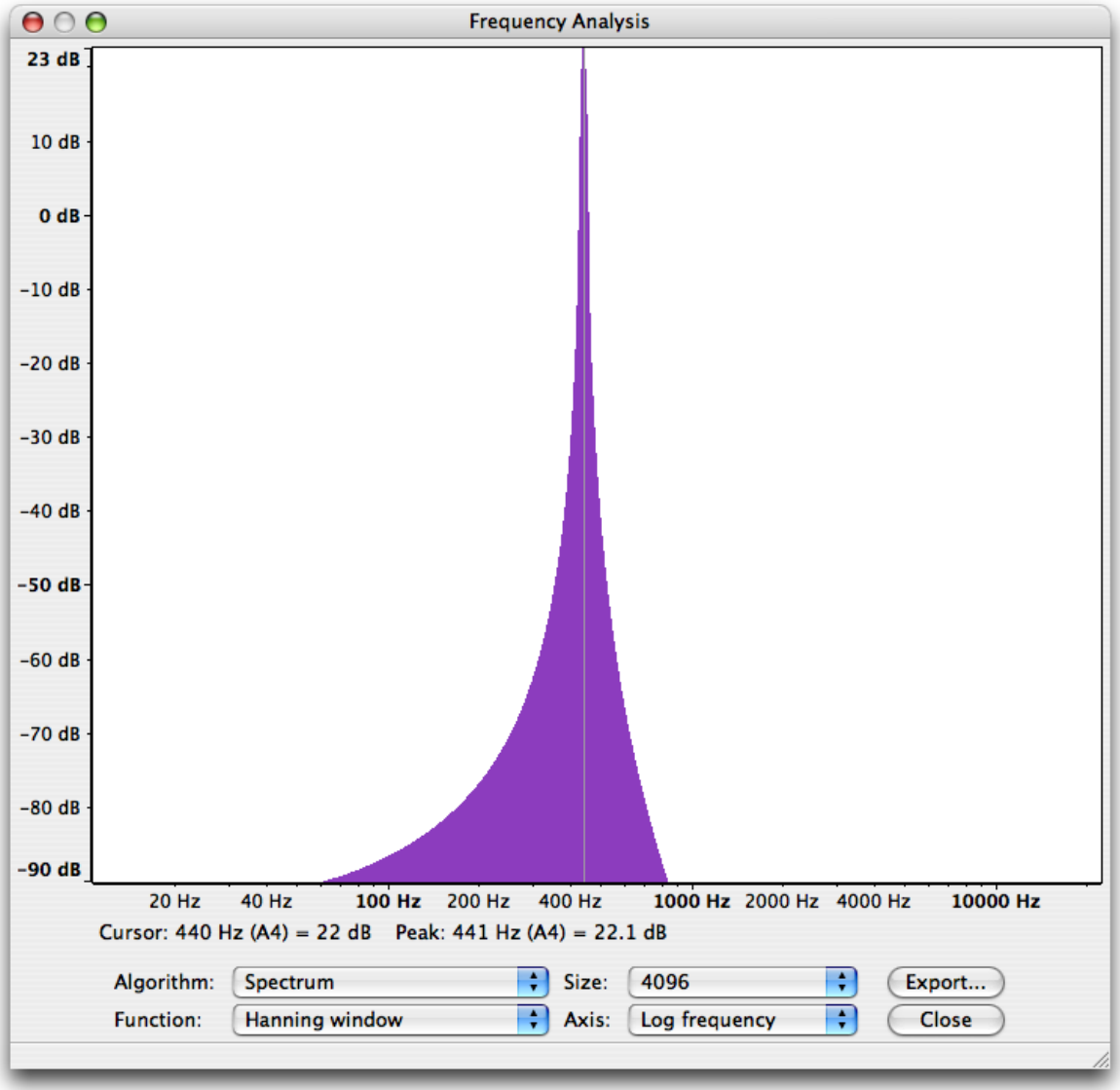
Sine wave at 440 Hz (Duration 100 ms): Play



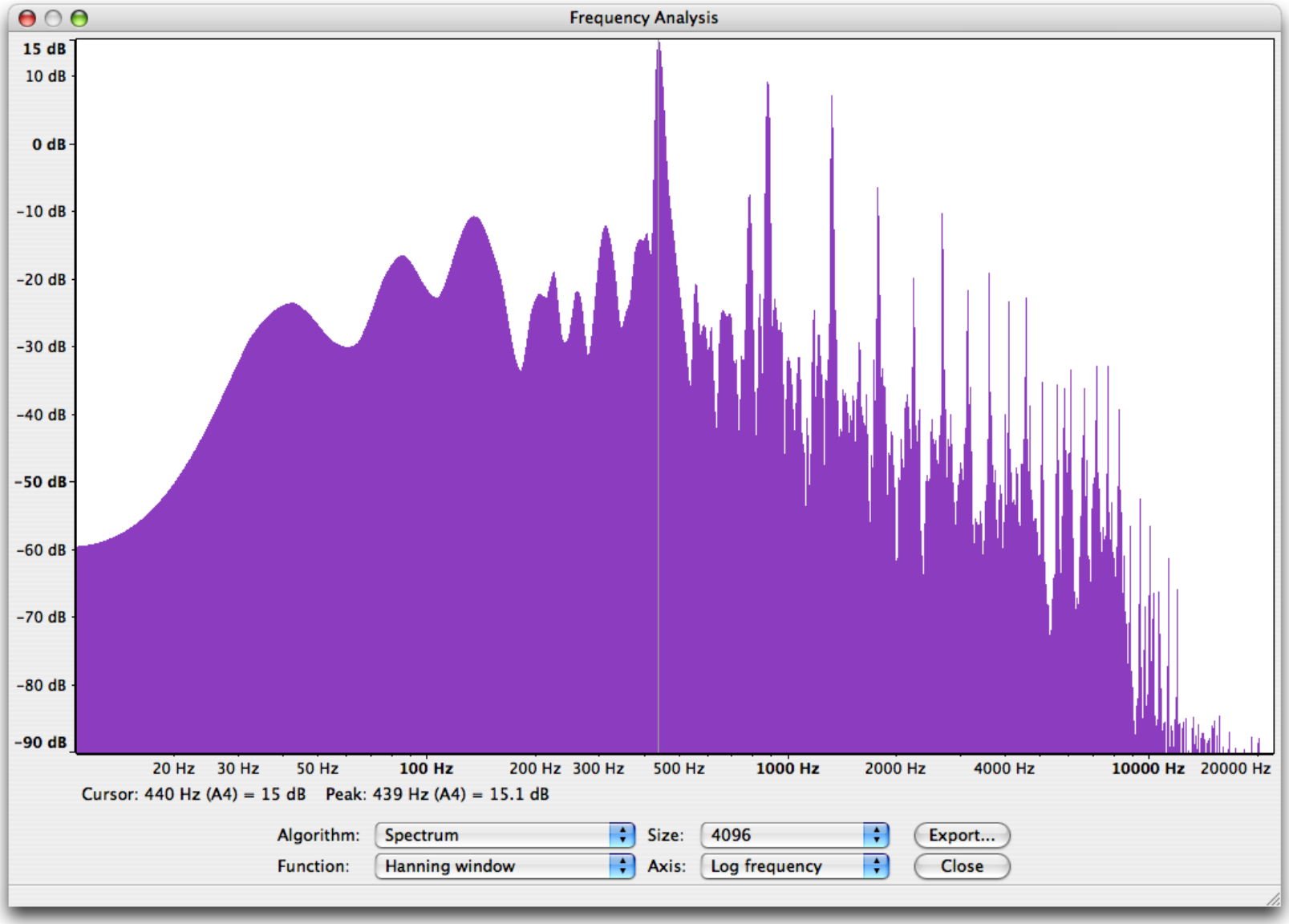
First 100ms of piano A4: Play



Spectrum of Sine Wave



Spectrum of Piano A4



Disclaimer :-)

I am not an expert in any of the following fields

- signal processing
- musicology
- statistics

Everything presented is to the best of my knowledge

Technical details sometimes omitted. Hopefully:

- Discussion about the overall picture
- Not too many mistakes

Similarity of Melodies

Klaus Frieler (2006):
Generalized N-gram Measures for Melodic Similarity.
In: *Data Science and Classification*, Springer, pp. 289–298.
http://dx.doi.org/10.1007/3-540-34416-0_31

Abstraction needed: Melody as Time Series

Definition

Let $j \leq k \in \mathbb{N}$. A sequence ϕ over some set $\mathbb{E}$ is a map

$$\phi : [j : k] \to \mathbb{E}$$
$$i \mapsto \phi_i \, .$$

ϕ is in normal form if $j = 0$.

Melody Abstraction

Definition

Let $\phi = (\phi_i)_{i \in [j:k]}$ be a sequence. An *n*-gram of ϕ is a subsequence $(\phi_i)_{i \in [l:m]}$ where $m - l + 1 = n$ and $i \leq l \leq m \leq k$.

Definition

Let $\mathbb{E}$ be an event space. A sequence $\mu : [0 : n-1] \to \mathbb{R} \times \mathbb{E}$, $i \mapsto \mu_i =: (t_i, p_i)$, is called *melody* if

$$i < j \Longrightarrow t_i < t_j \,.$$

Assumption here: Onset and pitch are sufficient to capture the "essence" of a melody

Similarity Measures (1/2)

Basic Choice about what to measure:

- Dissimilarity

  More common to mathematics (distance measures, metrics)

- Similarity

  Judging similarity far more familiar for musical experts

Definition

Let $\mathcal{M}$ be a set of melodies. A similarity map σ is a map $\sigma : \mathcal{M} \times \mathcal{M} \to [0, 1]$ such that the following is fulfilled:

- Symmetry: $\sigma(\mu, \mu') = \sigma(\mu', \mu)$
- Self-identity: $\sigma(\mu, \mu) = 1$ and $\sigma(\varepsilon, \mu) = 0$

(We denote by ε the empty melody.)

Similarity Measures (2/2)

Further requirement for melodies: Invariance under

- (Pitch) transposition,
- Time Shift,
- Tempo Change.

General idea in the following:

- Counting common n-grams

Further Preliminary Definitions

Let $\phi = (\phi_i)_{i \in [j:k]}$ be a sequence. Then:

- $|\cdot|$ is the length operator,
- $\widehat{\cdot}$ is the normalizing operator

such that $|\phi| = k - j + 1$ and $\widehat{\phi}_i = \phi_{i+j}$.

For any $n \in \mathbb{N}$:

$$\mathcal{S}_n(\phi) := \{r \mid r \text{ is } n\text{-gram of } \phi\} \quad \text{and}$$
$$\widehat{\mathcal{S}}_n(\phi) := \{\widehat{r} \mid r \text{ is } n\text{-gram of } \phi\}$$

The frequency of an n-gram r with respect to ϕ is defined as:

$$f_\phi(r) = |\{u \in \mathcal{S}_{|r|}(\phi) \mid \widehat{r} = \widehat{u}\}|$$

n-Gram Measures (1/2)

Definition

1. The Count-Distinct measure (CDM) is the count of n-grams common to both sequences:

$$S_n^{\mathrm{CDM}}(\phi, \varphi) := |\widehat{\mathcal{S}}_n(\phi) \cap \widehat{\mathcal{S}}_n(\varphi)|$$

2. The Sum-Common measure (SCM) is the sum of frequencies of n-grams common to both sequences:

$$S_n^{\mathrm{SCM}}(\phi, \varphi) := \sum_{r \in \widehat{\mathcal{S}}_n(\phi) \cap \widehat{\mathcal{S}}_n(\varphi)} \left[f_\phi(r) + f_\varphi(r) \right]$$

Normalization to get similarity measure:

$$\sigma_n^{\mathrm{CDM}}(\phi, \varphi) := \frac{S_n^{\mathrm{CDM}}(\phi, \varphi)}{\frac{1}{2} \cdot \left[|\widehat{\mathcal{S}}_n(\varphi)| + |\widehat{\mathcal{S}}_n(\phi)| \right]} \qquad \sigma_n^{\mathrm{SCM}}(\phi, \varphi) := \frac{S_n^{\mathrm{SCM}}(\phi, \varphi)}{|\phi| + |\varphi| - 2(n-1)}$$

n-Gram Measures (2/2)

Definition (continued)

3. The Ukkonen measure (UM) counts the absolute differences of frequencies of all distinct n-grams of both sequences:

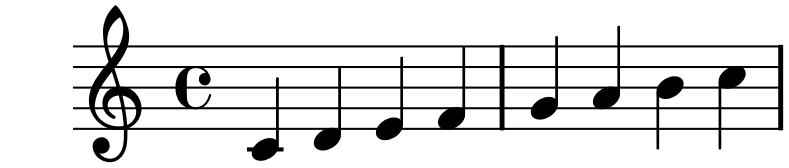
$$S_n^{\mathrm{UM}}(\phi, \varphi) := \sum_{r \in \widehat{\mathcal{S}}_n(\phi) \cup \widehat{\mathcal{S}}_n(\varphi)} |f_\phi(r) - f_\varphi(r)|$$

Transform into similarity measure:

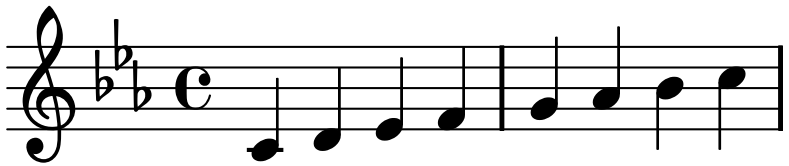
$$\sigma_n^{\mathrm{UM}}(\phi, \varphi) := 1 - \frac{S_n^{\mathrm{UM}}(\phi, \varphi)}{|\phi| + |\varphi| - 2(n-1)}$$

Example: Similarity of Melodies (1/2)

C major and C minor scale:



Play



Play

According to invariance w.r.t. transposition, tempo:

$$\phi = (2, 2, 1, 2, 2, 2, 1),\ \varphi = (2, 1, 2, 2, 1, 2, 2)$$

Example: Similarity of Melodies (2/2)

Two 6-grams for each melody:

$$r_1 = (2, 2, 1, 2, 2, 2), r_2 = (2, 1, 2, 2, 2, 1)$$
$$s_1 = (2, 1, 2, 2, 1, 2), s_2 = (1, 2, 2, 1, 2, 2)$$

We get $\sigma_n^\star(\phi, \varphi) = 0$ for all measures.

Consider only first 6 tones and 4-grams:

$$\phi = (2, 2, 1, 2, 2), \ \varphi = (2, 1, 2, 2, 1)$$

Two 4-grams for each melody:

$$r_1 = (2, 2, 1, 2), r_2 = (2, 1, 2, 2)$$
$$s_1 = (2, 1, 2, 2), s_2 = (1, 2, 2, 1)$$

We get $\sigma_n^\star(\phi, \varphi) = \frac{1}{2}$ for all measures.

Generalized Similarity Measures

Definitions of $S_n^\star(\phi, \varphi)$ can be rewritten in terms of frequencies:

$$S_n^{\mathrm{CDM}}(\phi, \varphi) := |\widehat{\mathcal{S}}_n(\phi) \cap \widehat{\mathcal{S}}_n(\varphi)| = \sum_{r \in \mathcal{S}_n(\phi)} 1_{\mathcal{S}_n(\varphi)}(r) \cdot \frac{1}{f_\phi(r)}$$

$$S_n^{\mathrm{SCM}}(\phi, \varphi) := \sum_{r \in \widehat{\mathcal{S}}_n(\phi) \cap \widehat{\mathcal{S}}_n(\varphi)} \left[f_\phi(r) + f_\varphi(r) \right]$$

$$= \sum_{r \in \mathcal{S}_n(\phi)} 1_{\mathcal{S}_n(\varphi)}(r) \cdot \left(1 + \frac{f_{\varphi(r)}}{f_\phi(r)} \right)$$

$$\vdots$$

Generalize the notion of frequency. For a similarity measure σ, let

$$\nu_\phi(r) := \sum_{u \in \mathcal{S}_{|r|}(\phi)} \sigma(\widehat{u}, \widehat{r}) \geq |\{u \in \mathcal{S}_{|r|}(\phi) \mid \widehat{r} = \widehat{u}\}| = f_\phi(r)\,.$$

Excursion: Edit-Distance

Edit-Distance $d(\phi, \varphi)$ between two sequences in normal form:

- Defined as number of deletion, insertion, and substitution steps

Typical example of dynamic programming:

	ε	S	a	t	u	r	d	a	y
ε	0	1	2	3	4	5	6	7	8
S	1	0	1	2	3	4	5	6	7
u	2	1	1	2	2	3	4	5	6
n	3	2	2	2	3	3	4	5	6
d	4	3	3	3	3	4	3	4	5
a	5	4	3	4	4	4	4	3	4
y	6	5	4	4	5	5	5	4	3

Recurrence relation for $i \in [0 : |\phi|]$, $j \in [0 : |\varphi|]$:

$$D(i,j) := \min \left\{ \begin{array}{l} D(i-1,j)+1, \\ D(i,j-1)+1, \\ D(i-1,j-1)+\delta(\phi_i, \varphi_j) \end{array} \right\} \qquad \begin{array}{l} D(i,0) := i \\ D(0,j) := j \end{array}$$

Summary and Criticism

Summary:

- Goal: Similarity of monophonic melodies
- Method: Melodies as time series
- Generalizing measures based on "n-grams", i.e., subsequences of fixed length

Criticism ($\rightarrow$ discussions? :-))

- Why only n-grams for arbitrary (but fixed) n's?
- Generalization unmotivated and "arbitrary"
- burdensome yet still imprecise mathematical formalisms (tried to mitigate here)

Evaluating Similarity Approaches

Daniel Müllensiefen, Klaus Frieler (2006):
Evaluating Different Approaches to Measuring the Similarity of Melodies.
In: *Data Science and Classification*, Springer, pp. 299–306.
http://dx.doi.org/10.1007/3-540-34416-0_32

Several studies to comparison similarity measurement for melodies:

- "Very different similarity values"
- "not clear which one is the most adequate"

General Idea of Similarity Algorithms

1. Basic transformations (representations)
   - Projections: E.g., to pitch component
   - Differentiation
2. Main transformations, e.g.:
   - Rhythmical weighting: Assume melody is quantized, replace pitch of duration $n \cdot T$, $n > 1$, by sequence of n tone with duration T
   - Contourization: Idea that perceptionally important notes are the extrema
     $\rightarrow$ substitute pitches in between by linear interpolation
3. Computation

Experimental Evaluation

Subjects: Musicology students with longtime musical experience

1st test: Similarity of 84 melody pairs with constructed errors

- Generated from 14 original "western popular" melodies
- Rhythm errors, Pitch errors, etc.

2nd and 3rd test: Also completely different melodies

Only let subjects with stable judgement continue:

- 1st test: 23 out of 82
- 2nd test: 12 out of 16
- 3rd test: 5 out of 10

Results

Homogeneity of human similarity judgement:

- High correlation between judgements of selected subjects
- Hypothesis: Objective similarity at least for "western" experts
- "Conceptual Foundation for statistical modeling"

Human notion of similarity is "adaptive":

- 1^{st} test with only slightly altered melodies: Pitch information sufficient
- 2^{nd} test: When melodies are different, subjects' ratings best modeled by including rhythmical information

"Unrelated melodies that differ strongly [...] are hard to relate". :-)

Estimate Pitch in General Audio Data

Katrin Sommer, Claus Weihs (2006):
Using MCMC as a Stochastic Optimization Procedure for Musical Time Series.
In: *Data Science and Classification*, Springer, pp. 307-314.
http://dx.doi.org/10.1007/3-540-34416-0_33

Pitch estimation of monophonic sound by joint estimation of overtones

- Based on a model by Davy and Godsill (2002)
- Estimating parameters of the model requires computing multi-dimensional integrals
- An MCMC (Monte Carlo Markov Chain) algorithm is used

Harmonic Model (1/2)

Basic model:

$$y_t = \sum_{h=1}^{H} a_{h,t} \cos(2\pi h f_0 t) + b_{h,t} \sin(2\pi h f_0 t)) + \varepsilon_t$$

where y_t is the instantaneous amplitude at time t.

Idea:

- Tone is composed out of harmonics from H partial tones
- First partial is fundamental frequency f_0
- Remaining $H - 1$ partials are called overtones
- Amplitudes of each partial tone are time dependent
- ε_t is model error

Harmonic Model (2/2)

Estimating the parameters of the former model requires computing a multidimensional integral of the form

$$\int_{\Omega} f(\theta, f_0, H, \sigma_\varepsilon^2) p(\theta, f_0, H, \sigma_\varepsilon^2 \mid y) \, \mathrm{d}\theta \, \mathrm{d}f_0 \, \mathrm{d}H \, \mathrm{d}\sigma_\varepsilon^2 \, .$$

Standard numerical techniques generally inaccurate or too slow

- Therefore, use a Metropolis-Hastings MCMC algorithm
- Sufficient for this talk: Generalization of Monte Carlo Integration

Mini-Excursion: Monte Carlo Integration

Wanted:

$$I = \int_0^1 f(\omega)\mathrm{d}\omega$$

Let $(U_i)_{i\in\mathbb{N}}$ be independent random variable with $U_i \sim U(0,1)$.
According to strong law of large numbers:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(U_i) = \mathrm{E}[f(U_i)] = I$$

The probabilistic error bound decreases as $1/\sqrt{n}$.

Mainly suited for multi-dimensional integrals (Dimension $= d$):

- Standard numerical approaches need exponentially many samples in d
- Monte Carlo approach: Error bound independent of d

Summary and Criticism

Goal:

- Estimate pitch from general audio signals
- Davy and Godsill (2002): General model very successful
- New: Introduction of "technical" stochastical optimization

Criticism:

- Hardly (if at all) understandable without looking at original paper by Davy and Godsill (2002)

Sound characteristics

Claus Weihs, Gero Szepannek, Uwe Ligges, Karsten Luebke, Nils Raabe (2006):
Local Models in Register Classification by Timbre.
In: *Data Science and Classification*, Springer, pp. 315–322.
http://dx.doi.org/10.1007/3-540-34416-0_34

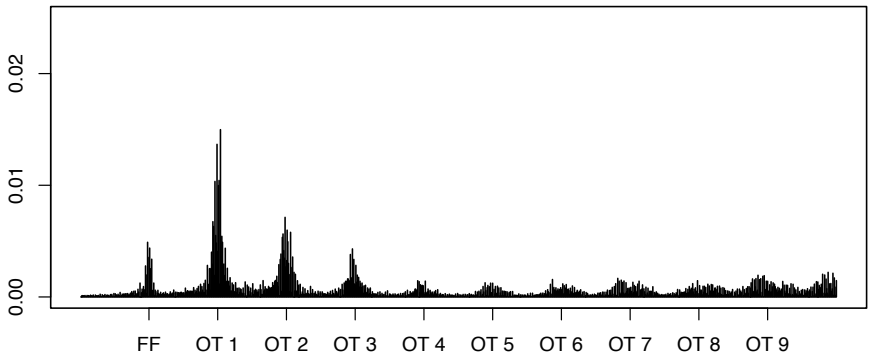
Follow-up paper to the following goal:

- Identify high and low musical register (Soprano, Alto vs. Tenor, Bass) by timbre
- After pitch information is eliminated from the spectrum

The Data

17 singers performing the song "Tochter Zion" (G.F. Händel)

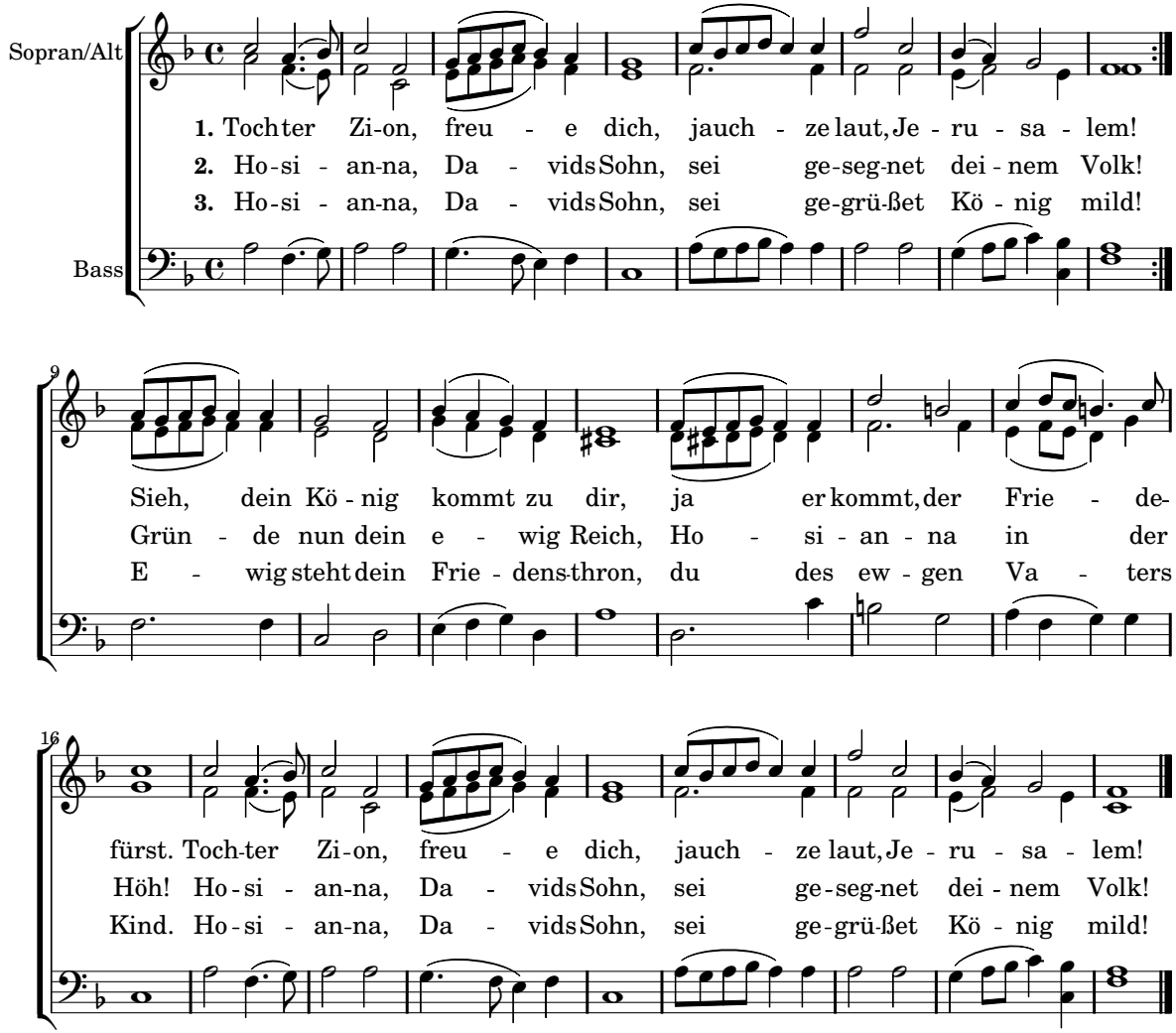
- Downsampled to 11,025Hz and standardized to interval $[-1, 1]$
- Analyses are based on characteristics derived from tones corresponding to single notes ($\rightarrow$ suitable segmentation)
- For identified notes: Derive pitch-idependent periodogram



- Consider only size and shape of first 13 partials. After cropping Fourier frequencies below some threshold
  - Mass: sum of the percentage shares
  - Width: difference between max and min frequency

Tochter Zion, freue Dich

Georg Friedrich Händel (1685-1759):



Classification

Common techniques: Linear Discriminant Analysis (LDA)

- $2 \cdot 13 = 26$ Variables are generated for every note separately
- Averaged over all notes
- Yields one single value of mass and width per harmonic and singer/instrument

Global Model:

- Training set includes samples from all instruments

Local Mode:

- Create own classification rules for each instrument
- Problem: Which classification rule to use?
  - Maximum-posterior rule: $\widehat{k} \in \arg\max_k (\max_l p_l(k \mid \mathbf{x}))$
  - Average-posterior rule: $\widehat{k} \in \arg\max_k \sum_l p_l(k \mid \mathbf{x})$
  - Majority Voting: ...

Results

Classification of voices alone gives good results. Explanations:

- Human mouth acts as a highpass filter
  - The lower the tone the less the mass of the fundamental frequency compared to 1^{st} overtone
  - Therefore: Sopranos have more mass in the fundamental f.
- Local model yields improvements when there are instruments
- Voice print of a professional bass singer: