Durham University

Introduction to Music Computing

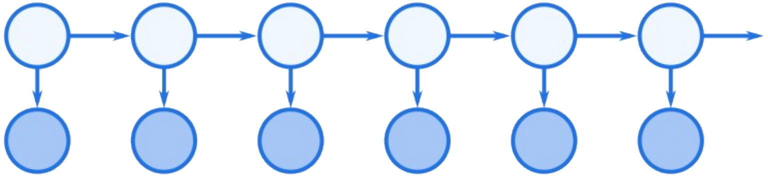
Sequential Models

n-gram, *k*-Markov, IDyOM

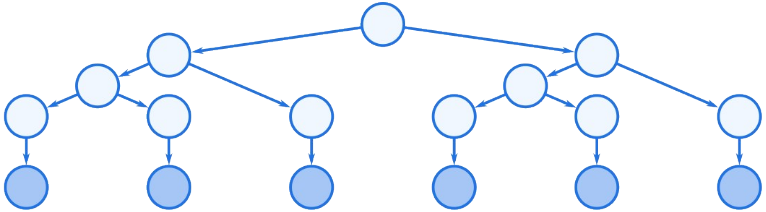
Dr Robert Lieck
robert.lieck@durham.ac.uk

Computational Models of Music

- **Sequential Models** (*n*-gram and (hidden) Markov models)



- **Hierarchical Models** (context-free grammars)



- **Neural Networks** (RNNs, Transformers, WaveNet)

Choosing a Model

Choosing a particular model or theory for describing the real-world implies:

- Making certain (simplifying) assumptions about the world
- Deciding what you are interested in capturing

Any model or theory:

- Has a particular scope of validity
  It is “correct” within that scope…
- Is tied to certain assumptions
  Which may or may not apply to your case…
- Looks at the world through a particular “lens”
  Which may or may not capture what you are interested in…

Music as a Sequence

Prélude No. 1 in C Major

from “Das Wohltemperierte Klavier” Book I
BWV 846

Johann Sebastian Bach
(1685 - 1750)

Music as a Sequence

- Conceptualise music as a sequence of events $e \in E$ from an alphabet/event space/domain E.
  - Could e.g. be notes, chords, harmonies.
- Try to predict the next event
  - What does it depend on? Previous events!
  - How many? 1, 2, 3, … ?

n-gram (or *k*-Markov) Models

The next event depends on the previous *n-1* (or *k*) events

- Event e_t at time t depends on events $e_{t-1}, \ldots, e_{t-(n-1)}$ (called the context).
- Tuple $(e_{t-(n-1)}, \ldots, e_{t-1}, e_t) = e_{t-(n-1)}^{t}$ is called an *n*-gram.
  (n=1: unigram, n=2: bigram).
- We want to know $p(e_t \mid e_{t-(n-1)}^{t-1})$, that is
  - the probability of e_t
  - given the previous n-1 events $e_{t-(n-1)}^{t-1}$
- Probabilities are proportional to *n*-gram counts.

Durham University

Naïve n-gram Model

- Count all n-grams in the data.
- Compute probabilities as relative counts
  → corresponds to a maximum likelihood estimate

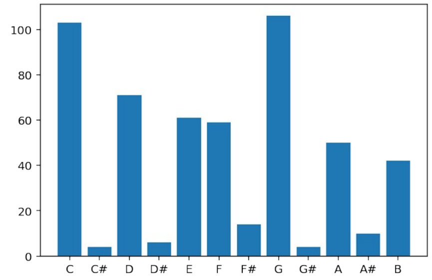
$$\boxed{p_n(e_t \mid e_{t-(n-1)}^{t-1}) = \frac{\#(e_{t-(n-1)}^{t})}{\sum_{\bar{e} \in E} \#(\bar{e}, e_{t-(n-1)}^{t-1})}}$$

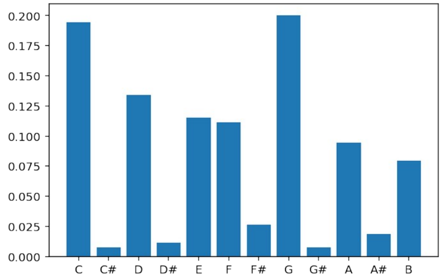
Durham University

Example: C Major Prelude

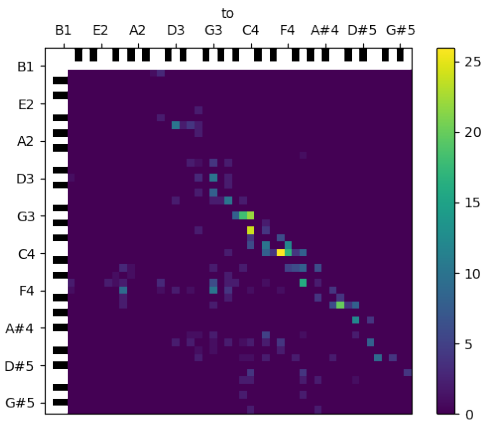
Counts | **Probabilities**

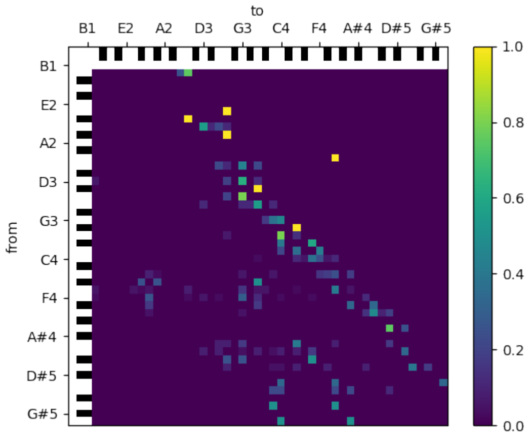
Unigrams (pitch classes)



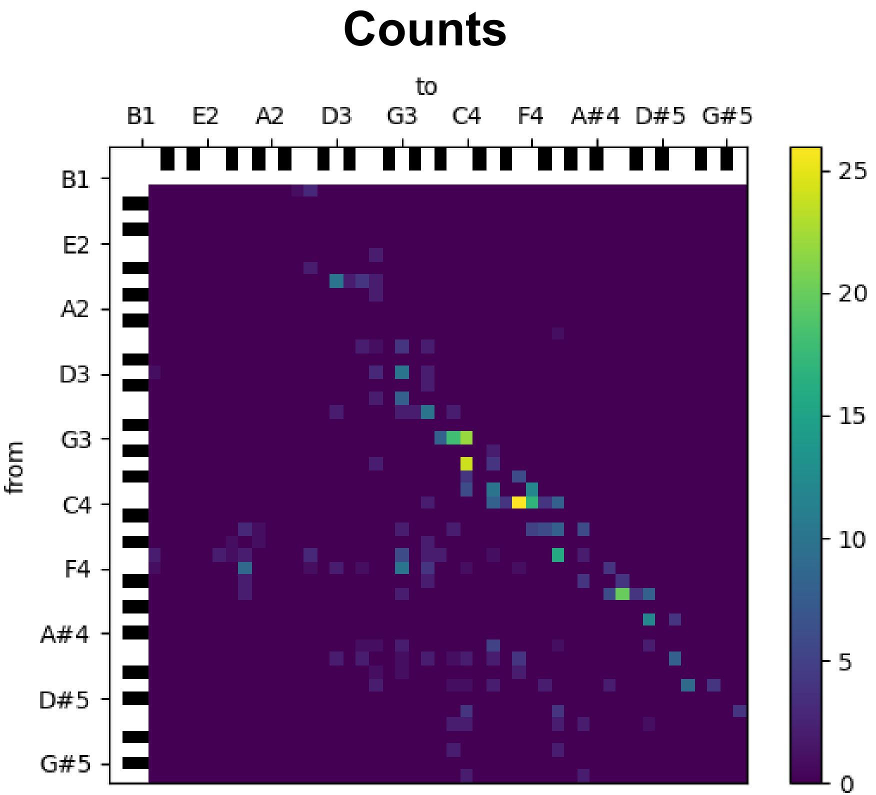


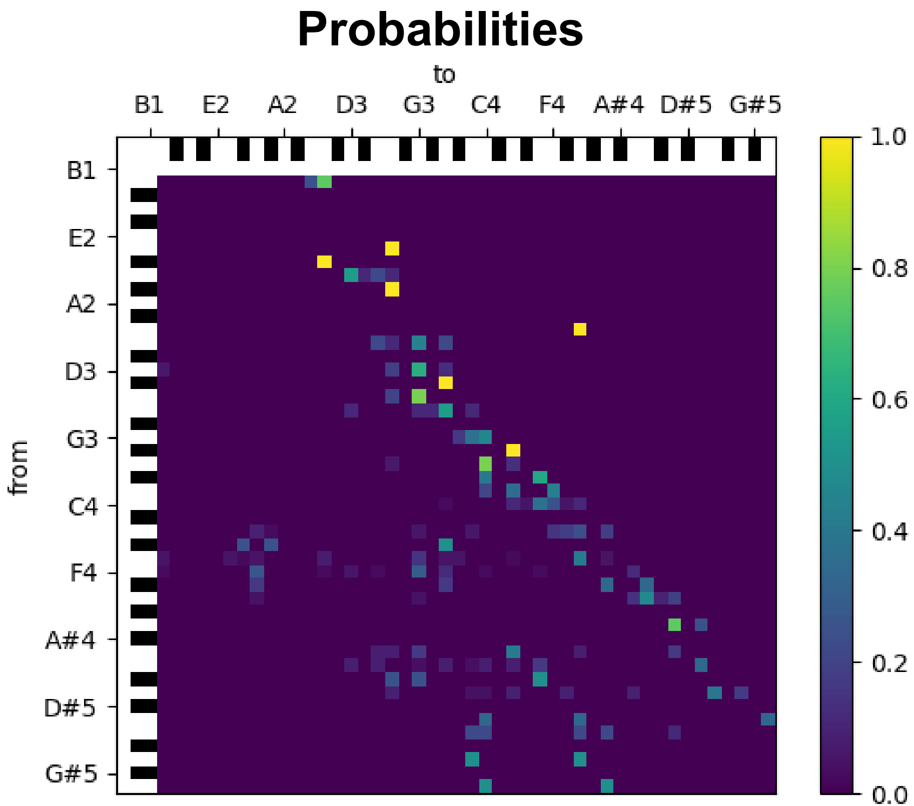
Bigrams (pitches)





Example: C Major Prelude (Bigrams)

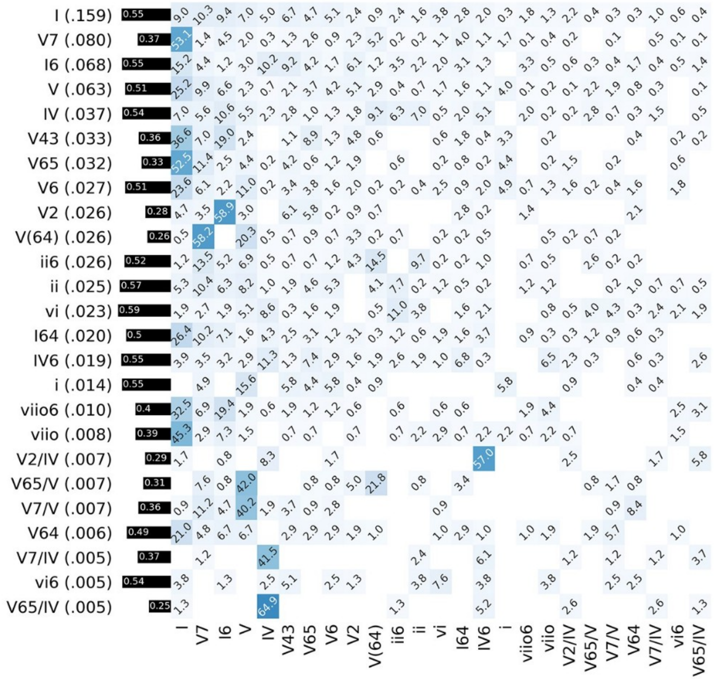




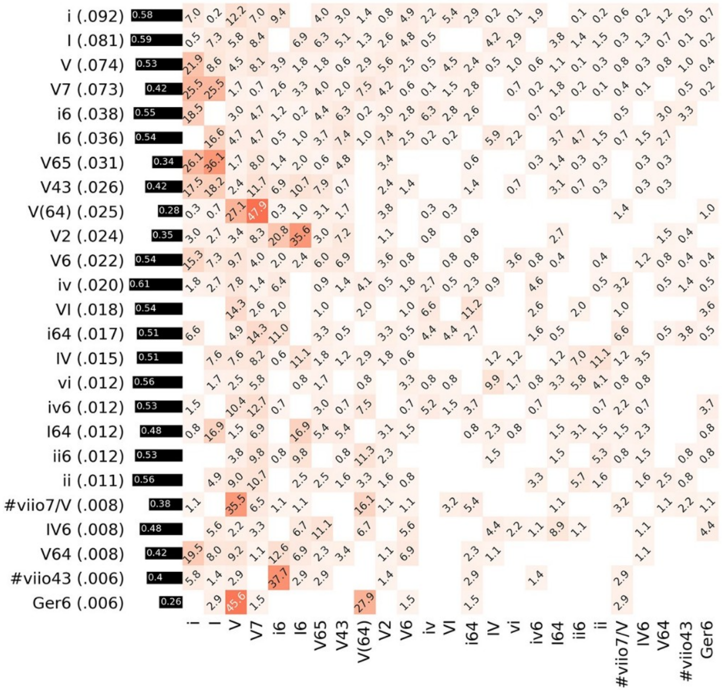
Durham University

Example: Beethoven’s String Quartets
(Chord Transitions)

Major



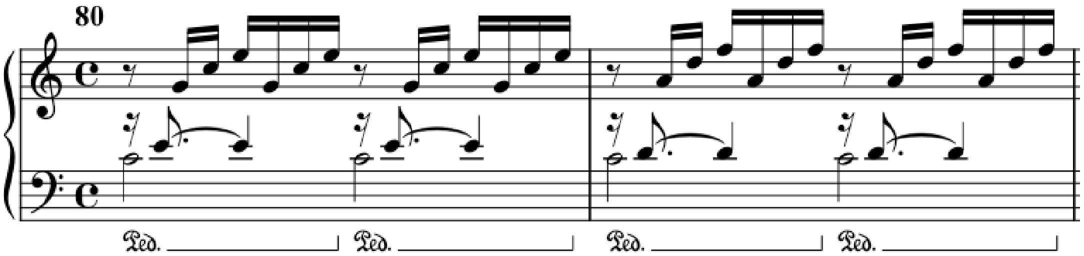
Minor



Moss FC, Neuwirth M, Harasim D, Rohrmeier M (2019) Statistical characteristics of tonal harmony: A corpus study of Beethoven’s string quartets. PLoS ONE

Naïve n-gram Model: Problems

$$p_n(e_t \mid e_{t-(n-1)}^{t-1}) = \frac{\#(e_{t-(n-1)}^{t})}{\sum_{\bar{e} \in E} \#(\bar{e}, e_{t-(n-1)}^{t-1})}$$



Naïve *n*-gram Model: Problems

- **Zero counts**
  → How to deal with unknown contexts?

- **Variable length *n***
  → How to combing different context lengths?

- **Different event spaces / viewpoints**
  → How to use e.g. interval or contour information to model pitch?

- **Short-term (online) versus long-term (offline) model**
  → How to model e.g. motifs versus style?

Zero Counts: Prior Counts

Add prior counts α > 0 to all events

- All probabilities are non-zero (like having seen everything α times before looking at the actual data)
- Limit α → 0: uniform distribution for unknown contexts
- Connection to Bayesian inference:
  - α is concentration parameter of Dirichlet prior
  - # are sufficient statistics of data
  - (# + α) gives the posterior distribution

$$\boxed{p_n(e_t \mid e_{t-(n-1)}^{t-1}) = \frac{\#(e_{t-(n-1)}^{t}) + \alpha}{\sum_{\bar{e} \in E} \#(\bar{e}, e_{t-(n-1)}^{t-1}) + \alpha}}$$

Durham University

Variable length *n*: Backoff

Choose *n* on the fly

- Start with long contexts (large *n*)
- Backoff to shorter contexts (smaller *n*)
- Recursion always terminates (unigram or uniform distribution)

$$p_n^{\text{backoff}}(e_t \mid e_{t-(n-1)}^{t-1}) = \begin{cases} p_n(e_t \mid e_{t-(n-1)}^{t-1}) & \text{if} \quad \beta(e_{t-(n-1)}^{t-1}) \\ p_{(n-1)}^{\text{backoff}}(e_t \mid e_{t-(n-2)}^{t-1}) & \text{else} \end{cases}$$

- β may depend on context and length *n*
- E.g. to avoid zero counts: $\beta(e_{t-(n-1)}^{t-1}) = \#(e_{t-(n-1)}^{t-1}) > 0$

Durham University

Variable length *n*: Smoothing

Linear combination of different context lengths *n*

- Give different weight to different contexts lengths
- Give more weight to longer contexts
- More general than backoff

$$\boxed{\begin{aligned} p_n^{\text{smooth}}(e_t \mid e_{t-(n-1)}^{t-1}) = \lambda(e_{t-(n-1)}^{t-1})\ p_n(e_t \mid e_{t-(n-1)}^{t-1}) + \\ \left(1 - \lambda(e_{t-(n-1)}^{t-1})\right)\ p_{(n-1)}^{\text{smooth}}(e_t \mid e_{t-(n-2)}^{t-1}) \end{aligned}}$$

- λ may depend on context and length *n*
- Equivalent to backoff for: $\lambda(e_{t-(n-1)}^{t-1}) = \begin{cases} 1 & \text{if} \quad \beta(e_{t-(n-1)}^{t-1}) \\ 0 & \text{else} \end{cases}$

Durham University

Multiple-Viewpoint Systems

Multiple views (features) of the data can be combined for preditions

To evaluate model, first map prediction to respective viewpoint

- **Basic:** Raw features; defined everywhere
- **Derived:** Computed from *basic*; defined if input defined
- **Linked:** Product/combination of other features
- **Threaded:** (Combination with) boolean feature; defined only at specific points

Basic:	Pitch	67	67	71	69	67	69	69	71
	Duration	48	48	48	48	48	48	48	48
Derived:	Interval	$\perp$	0	4	-2	-2	2	0	2
	ScaleDegree	0	0	4	2	0	2	2	4
Threaded:	ScaleDegree $\ominus$ FiB	$\perp$	0	$\perp$	$\perp$	$\perp$	2	$\perp$	$\perp$
Linked:	DurRatio $\otimes$ Interval	$\perp$	$\langle 1, 0\rangle$	$\langle 1, 4\rangle$	$\langle 1, -2\rangle$	$\langle 1, -2\rangle$	$\langle 1, 2\rangle$	$\langle 1, 0\rangle$	$\langle 1, 2\rangle$

Durham University

Combining Multiple Models

Arithmetic Mean

$$p^{\text{arith.}}(e) = \frac{\sum_i w_i \, p_i(e)}{\sum_i w_i}$$

Geometric Mean

$$p^{\text{geom.}}(e) = \frac{1}{Z}\left(\prod_i p_i(e)^{w_i}\right)^{\frac{1}{\sum_i w_i}} = \frac{1}{Z}\exp\frac{\sum_i w_i \, \log p_i(e)}{\sum_i w_i}$$

Weights w_i are hyper parameters and can either be optimised or heuristically chosen based on entropy of p_i.

Durham University

Long-Term and Short-Term Models

Long-Term Model

- Trained **offline** on a corpus of data
- Does not change during generation
- Captures style-specific characteristics

Short-Term Model

- Trained **online** while generating data
- Picks up on patterns in the data
- Captures piece-specific, motivic characteristics

Combined in the same way as multiple-viewpoint models!

References

1) Conklin D, Witten IH (1995) Multiple viewpoint systems for music prediction. Journal of New Music Research 24:51–73
2) Chen SF, Goodman J (1999) An empirical study of smoothing techniques for language modeling. Computer Speech & Language 13:359–394
3) Pearce M, Conklin D, Wiggins G (2004) Methods for combining statistical models of music. In: International Symposium on Computer Music Modeling and Retrieval. Springer, pp 295–312
4) Pearce MT (2005) The construction and evaluation of statistical models of melodic structure in music perception and composition. City University London
5) Whorley RP, Wiggins GA, Rhodes C, Pearce MT (2013) Multiple Viewpoint Systems: Time Complexity and the Construction of Domains for Complex Musical Viewpoints in the Harmonization Problem. Journal of New Music Research 42:237–266. https://doi.org/10.1080/09298215.2013.831457
6) Moss FC, Neuwirth M, Harasim D, Rohrmeier M (2019) Statistical characteristics of tonal harmony: A corpus study of Beethoven’s string quartets. PLoS ONE https://doi.org/10.1371/journal.pone.0217242