Multiple pitch transcription and melody harmonization with probabilistic musicological models

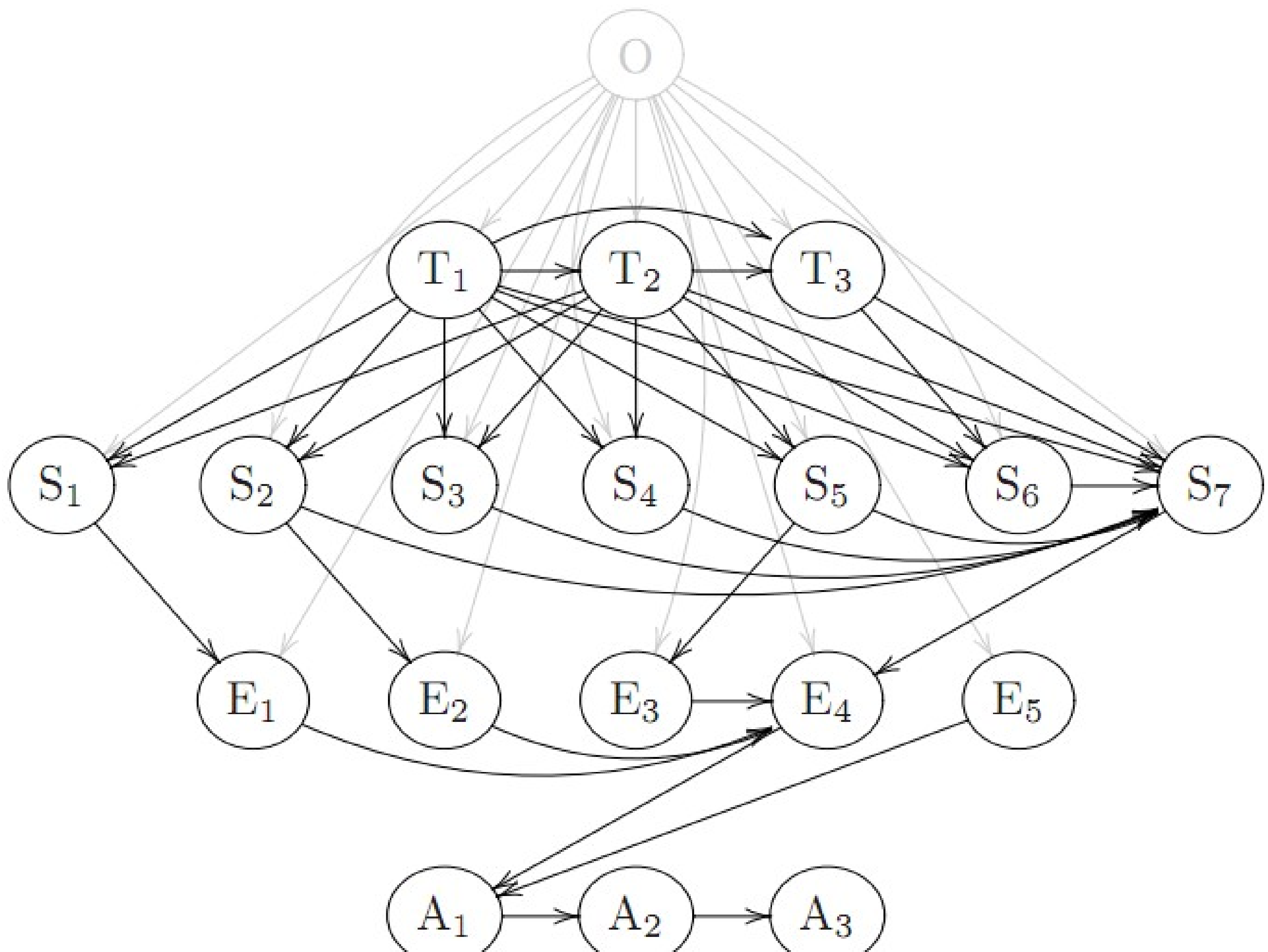
Stanisław A. Raczyński, Emmanuel Vincent

Introduction

- Musical quantities can be thought of as interconnected variables.
- Each variable holds information about itself and about others as well.
- For example, certain *chords* and their *progressions* (*e.g.*, ii-m7 | V-7 | I-maj7) will suggest that the genre is *jazz*, which will also implicate the use of certain *instruments* (*e.g.*, saxophone, piano, double bass)

Bayesian networks

- Relations between variables can be represented in a form of a Bayesian network [1]:

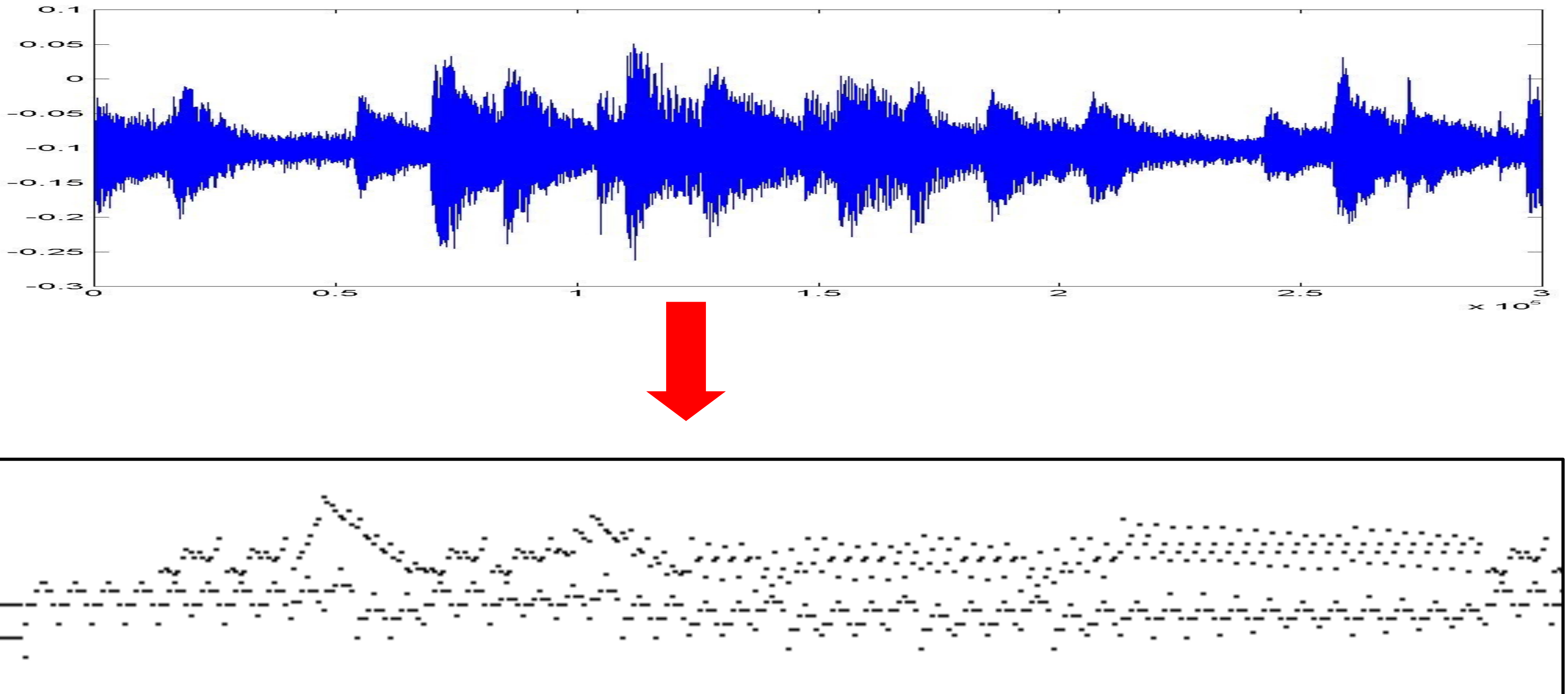


Relation to language models

- In processing natural language (*e.g.*, continuous speech recognition), probabilistic models of language are used and they are called *linguistic models* or *language models*.
- In music information retrieval, their equivalents are referred to as *musicological models* or *music models*.

Multiple pitch estimation

Estimating note *pitches*, *onsets* and *durations* given an audio recording



Current approaches

- The most popular approaches are based on Nonnegative Matrix Factorization (NMF).
- A spectrogram (typically obtained using a constant-Q or ERB filter bank) $\boldsymbol{X}$ of the recording is factorized to obtain the *dictionary matrix* $\boldsymbol{A}$ and the *salience matrix* $\boldsymbol{S}$:

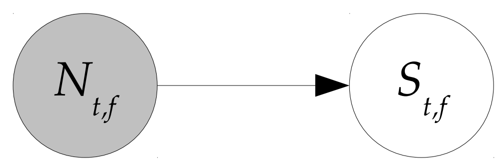
$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$$

- The salience matrix is then analyzed to find the positions of notes

Current approaches

- NMF is a mid-level representation of the audio.
- Typically, the salience values are analyzes individually, *e.g.*, thresholded.

Latent note variable

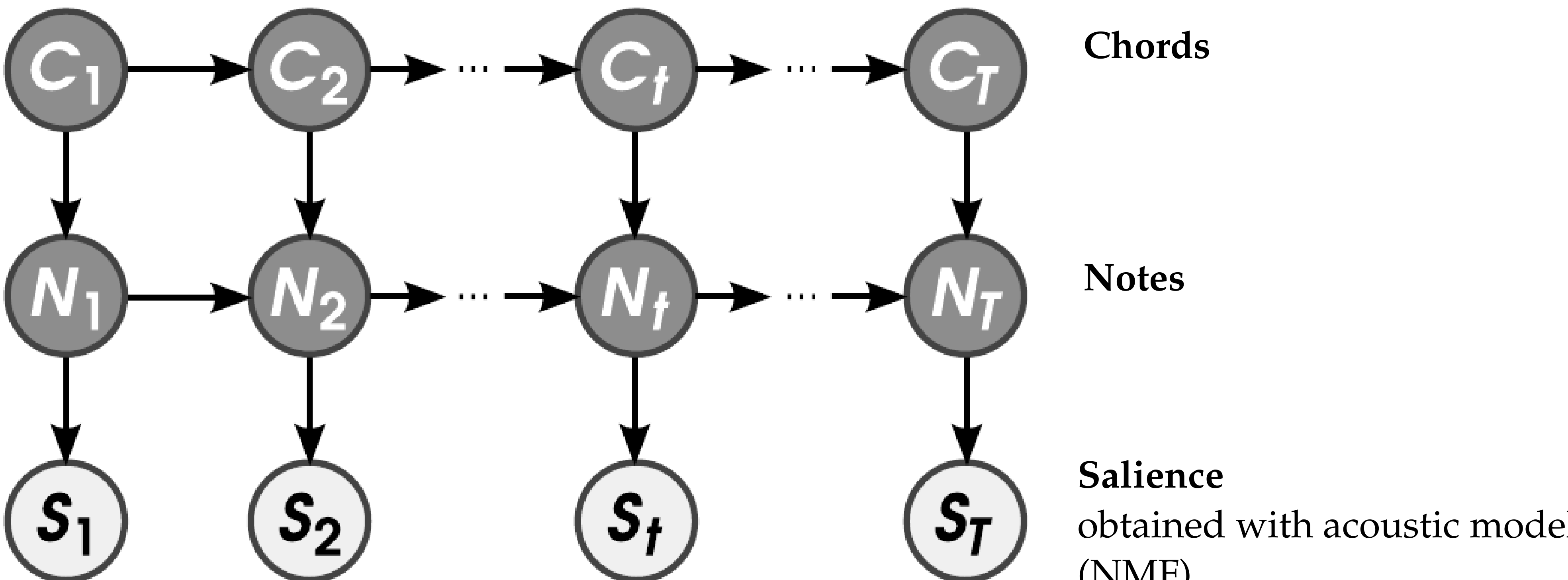


Observed salience variable

- Better results can be obtained if relations between the underlying binary note variables and more aspects of the music are modeled jointly.

Music pitch model

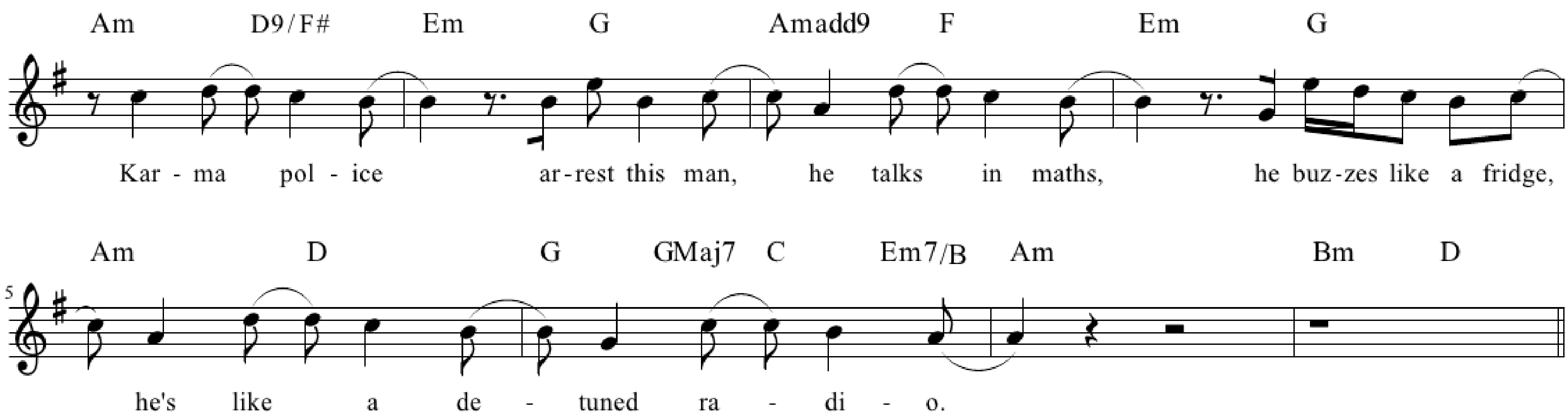
In our experiments we have used a Dynamic Bayesian Network to model relations between the latent and observed variables:



$$\mathrm{P}(\mathbf{N}) = \sum_{\mathbf{C}} \mathrm{P}(C_1)\mathrm{P}(\mathbf{N}_1|C_1) \cdot \prod_{t=2}^{T} \mathrm{P}(\mathbf{N}_t|\mathbf{N}_{t-1}, C_t)\mathrm{P}(C_t|C_{t-1})$$

Harmonization

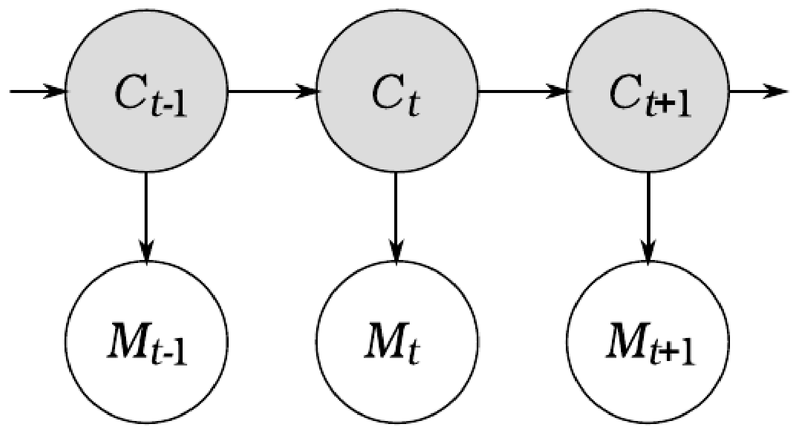
- Guessing the underlying *chord sequence* given a *melody*



- Used for automatic music composition, automatic accompaniment, *etc*.

Current approaches

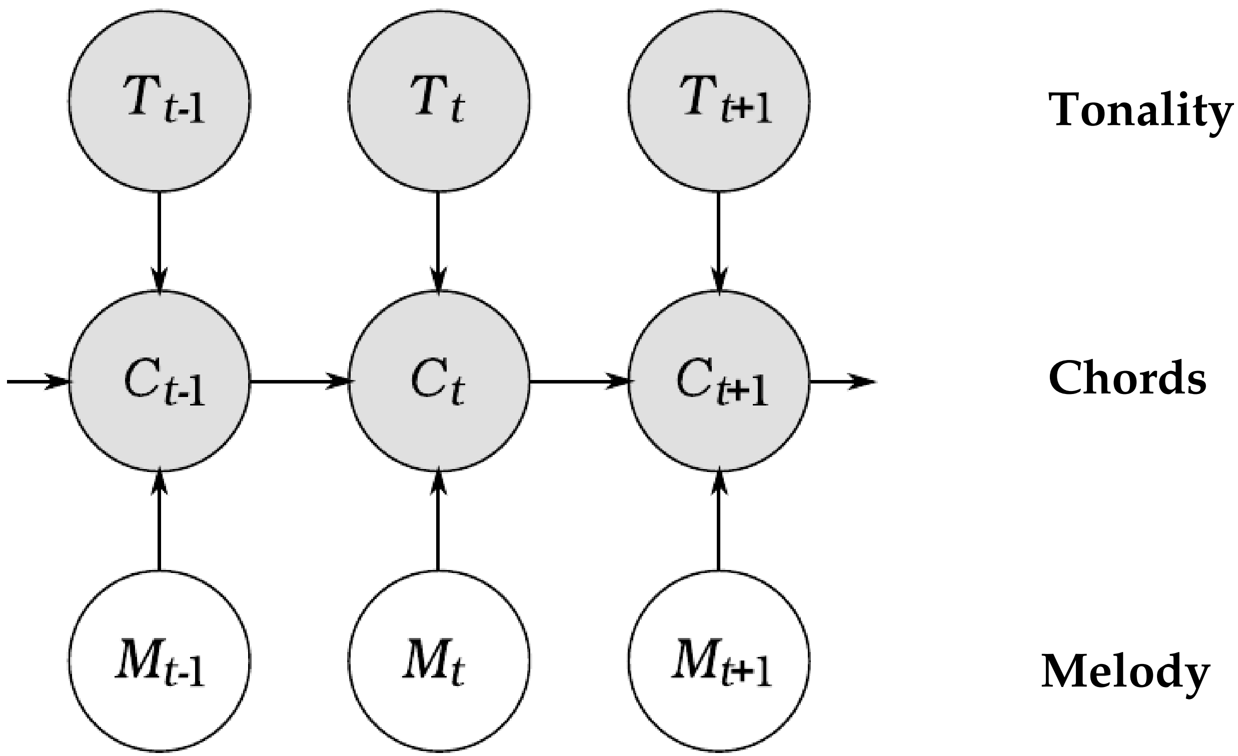
- A typical approach for harmonization uses Hidden Markov Models (HMMs) to model relations between the latent chords and the melody:



- This approach is used in such commercial applications as MySong [2] or Band-in-a-box [3].

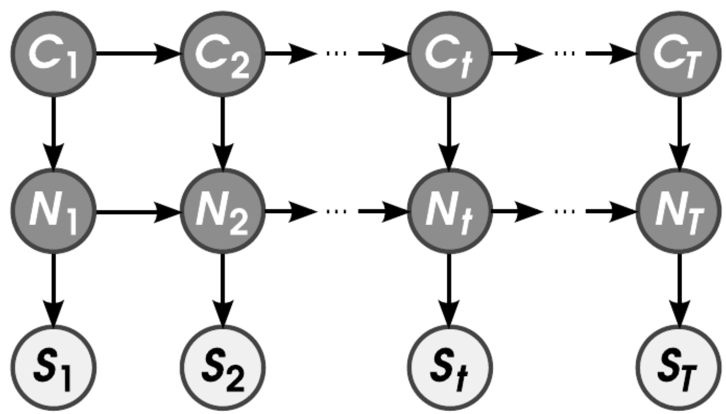
Music melody model

In our experiments we have used a Dynamic Bayesian Network to model relations between the latent and observed variables:



Model complexity

- Jointly modeling multiple variables causes the number of parameters to explode



$$\mathrm{P}(\mathbf{N}) = \sum_{\mathbf{C}} \mathrm{P}(C_1)\mathrm{P}(\mathbf{N}_1|C_1) \prod_{t=2}^{T} \mathrm{P}(\mathbf{N}_t|\mathbf{N}_{t-1}, C_t)\mathrm{P}(C_t|C_{t-1})$$

$2^K \times 24 \times 2^K =$ **$2.3 \cdot 10^{54}$** parameters for $K = 88$

Model interpolation

- Complexity can be reduced by approximating the joint model with a combination of simpler models – *model interpolation*.
- Model interpolation has been successfully used in natural language processing by Klakow [4].
- This technique is also used to reduce overfitting: models of different order are combined (*model smoothing*).

Model interpolation: linear

$$\mathrm{P}(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}) = \prod_{k=1}^{K} \mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1})$$

$$\mathrm{P}(N_{t,k}|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) \approx Z^{-1} \sum_i \lambda_i \mathrm{P}_i(N_{t,k}|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1})$$

$$Z = \sum_{l=0}^{1} \sum_i \lambda_i \mathrm{P}_i(N_{t,k} = l|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1})$$

Submodels P_i use only a small subset of the conditioning variable set, *e.g.*:

$$\mathrm{P}_2(N_{t,k}|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|N_{t-1,k})$$

Model interpolation: log-linear

$$\mathrm{P}(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}) = \prod_{k=1}^{K} \mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1})$$

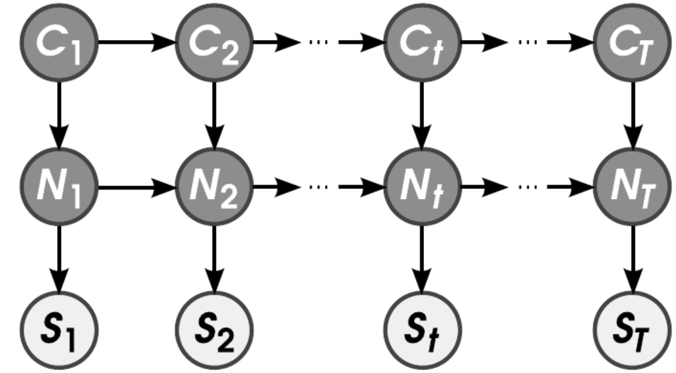
$$\mathrm{P}(N_{t,k}|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) \approx Z^{-1} \prod_{i} \mathrm{P}_i(N_{t,k}|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1})^{\lambda_i}$$

$$Z = \sum_{l=0}^{1} \prod_{i} \mathrm{P}_i(N_{t,k} = l|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1})^{\lambda_i}$$

Models and submodels used

and their trained parameter values

Pitch submodels



Harmony $$\mathrm{P}_1(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(\,\mathrm{inter}\{k; \mathrm{root}\{C_t\}\}|\mathrm{mode}\{C_t\})$$

Duration $$\mathrm{P}_2(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|N_{t-1,k})$$

Voice $$\mathrm{P}_3(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|M_{t,k})$$
$$M_{t,k} = |k - j|$$

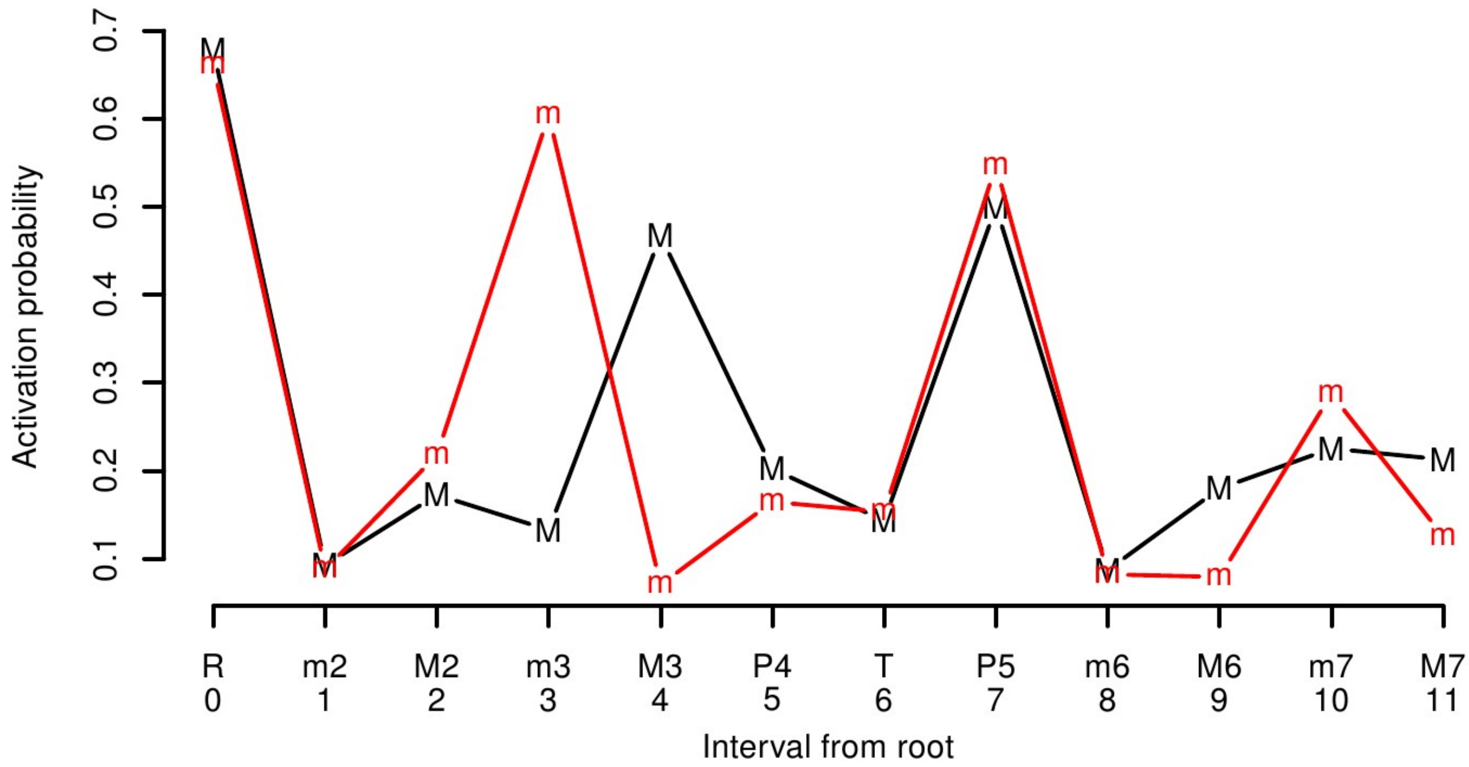
Polyphony $$\mathrm{P}_4(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|L_{t,k})$$
$$L_{t,k} = \sum_{m=1}^{k-1} N_{t,m}$$

Neighbor $$\mathrm{P}_5(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|N_{t,k-1}, N_{t,k-2})$$

Harmony submodel

- Independent of octave, depends only on the chord *mode* and the *interval* from chord's root:

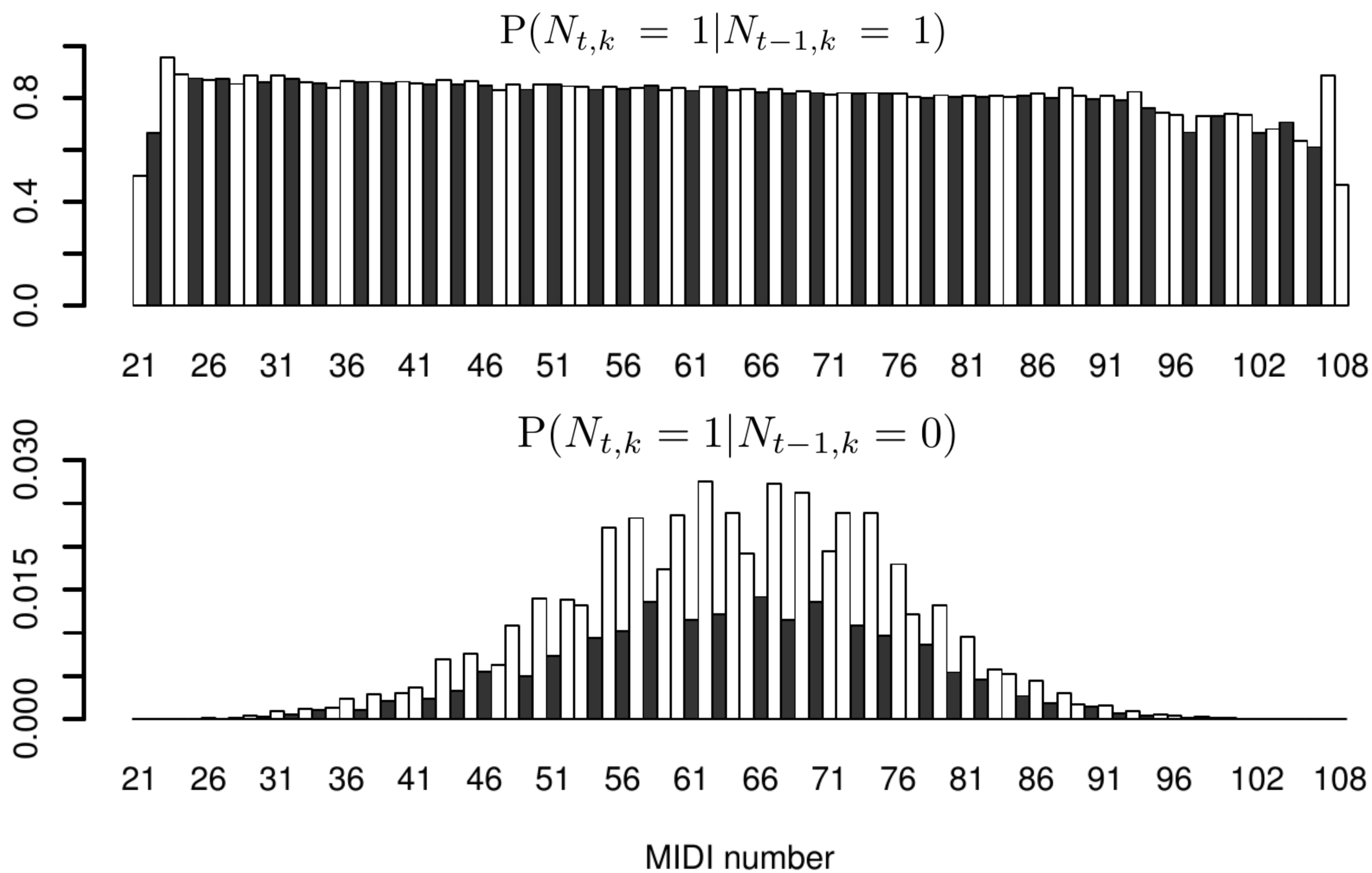
$$P_1(\mathbf{N}_t | C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = P(\text{ inter}\{k; \text{root}\{C_t\}\} | \text{mode}\{C_t\})$$



Duration submodel

- Simple binary bigram model:

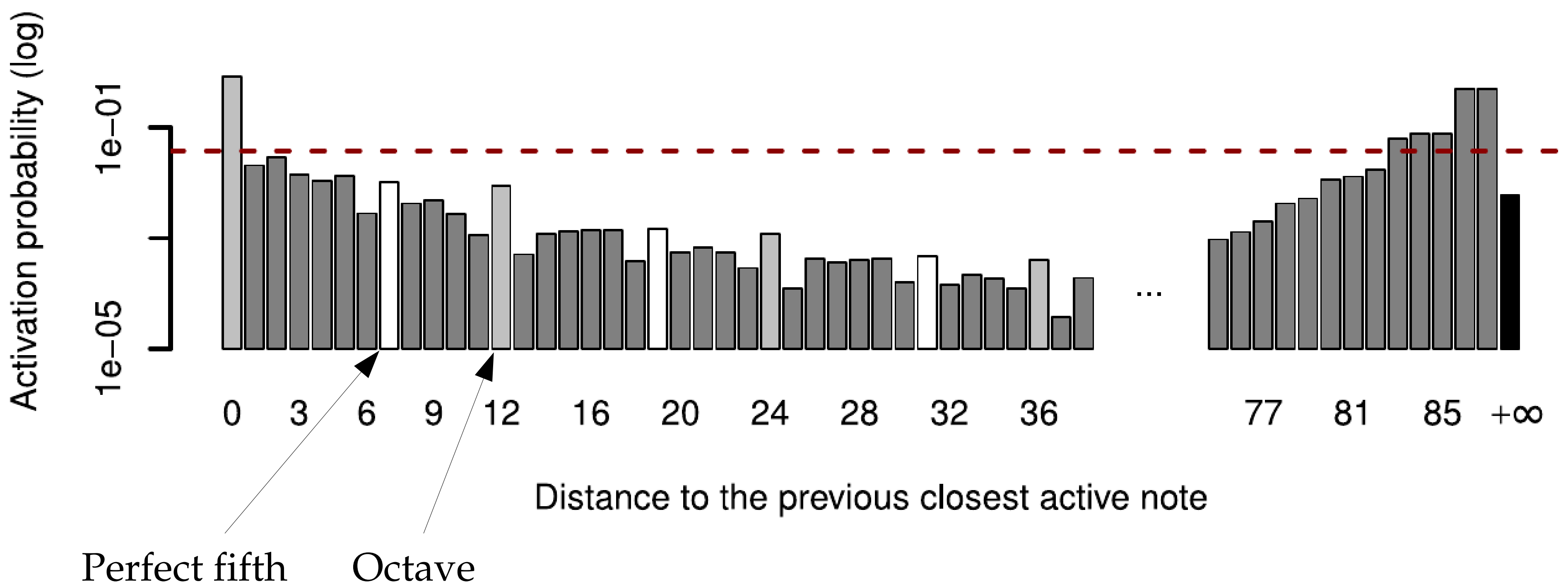
$$\mathrm{P}_2(N_{t,k}|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|N_{t-1,k})$$



Voice submodel

- Pitch activity depends only on the distance to the closest active pitch in the previous frame:

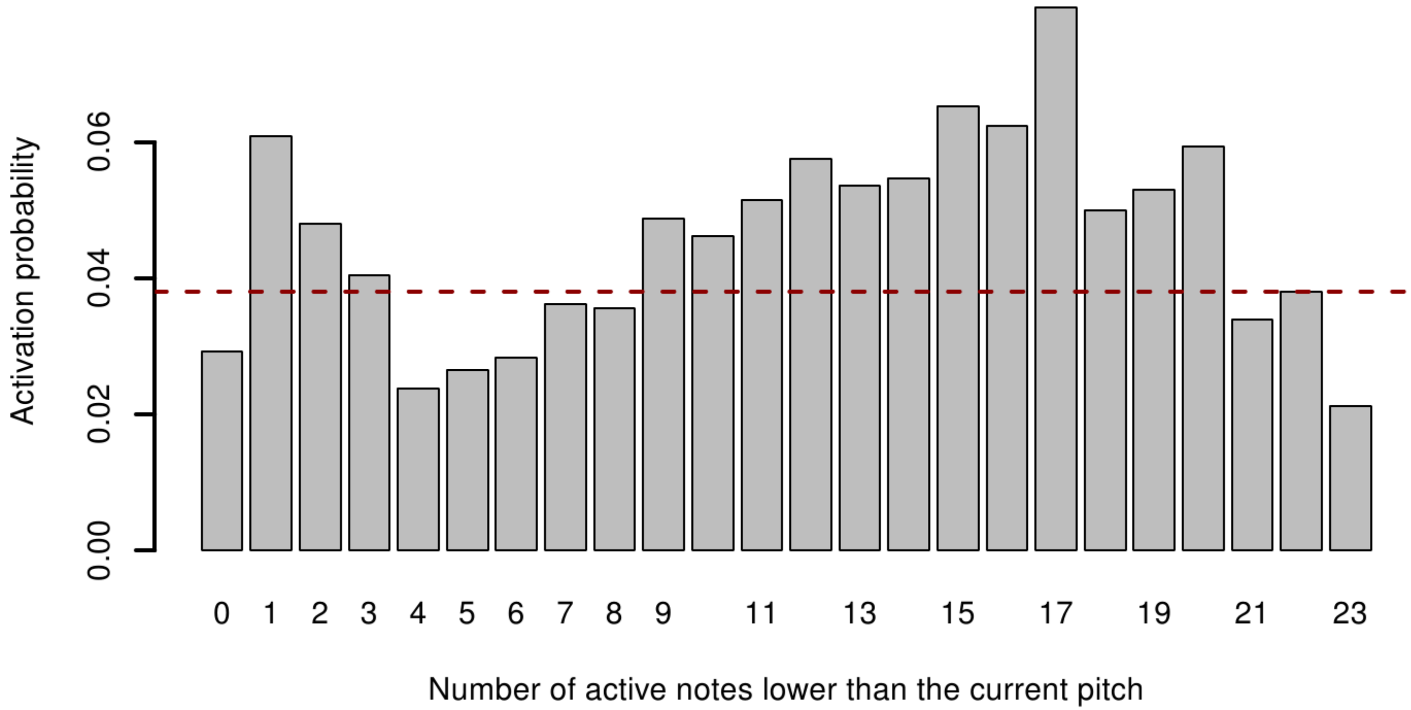
$$P_3(\mathbf{N}_t | C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = P(N_{t,k} | M_{t,k})$$



Polyphony submodel

- Pitch activity depends only on the number of active notes below the current pitch:

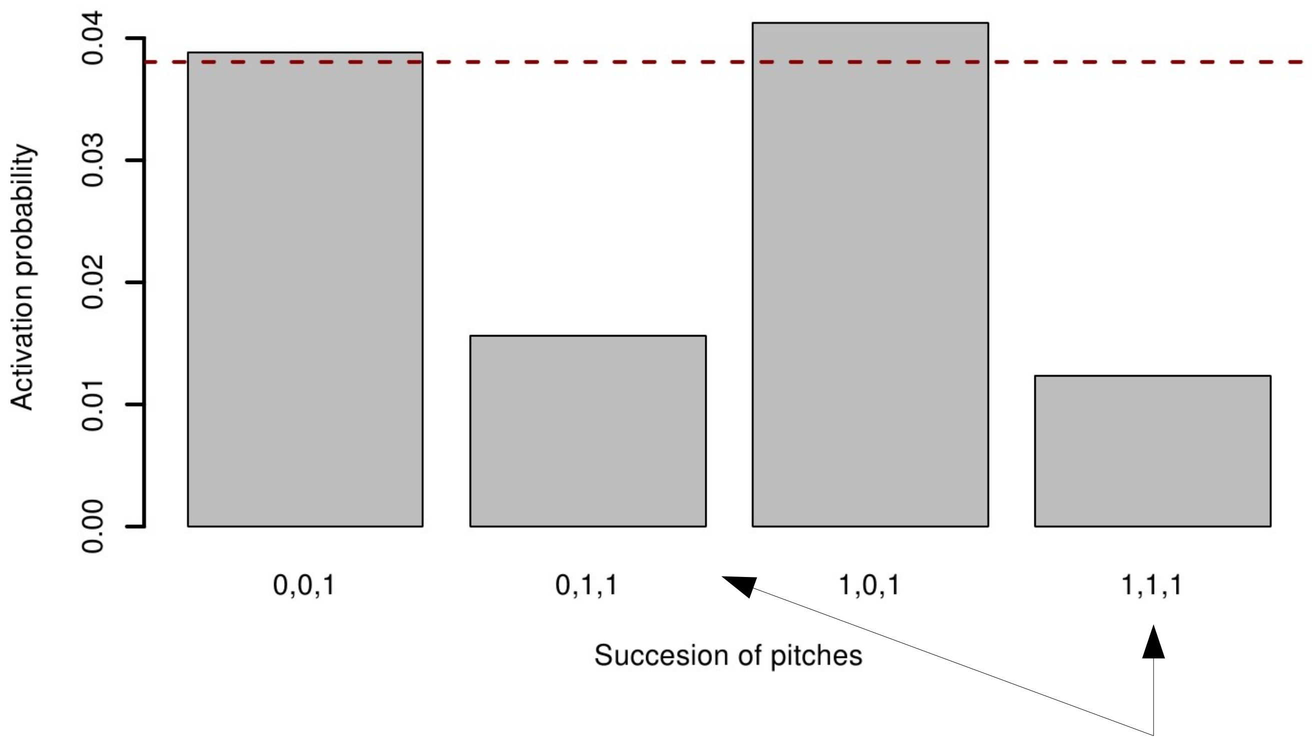
$$P_4(\mathbf{N}_t | C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = P(N_{t,k} | L_{t,k})$$



Neighbor submodel

- A binary trigram model in the frequency domain:

$$\mathrm{P}_5(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|N_{t,k-1}, N_{t,k-2})$$

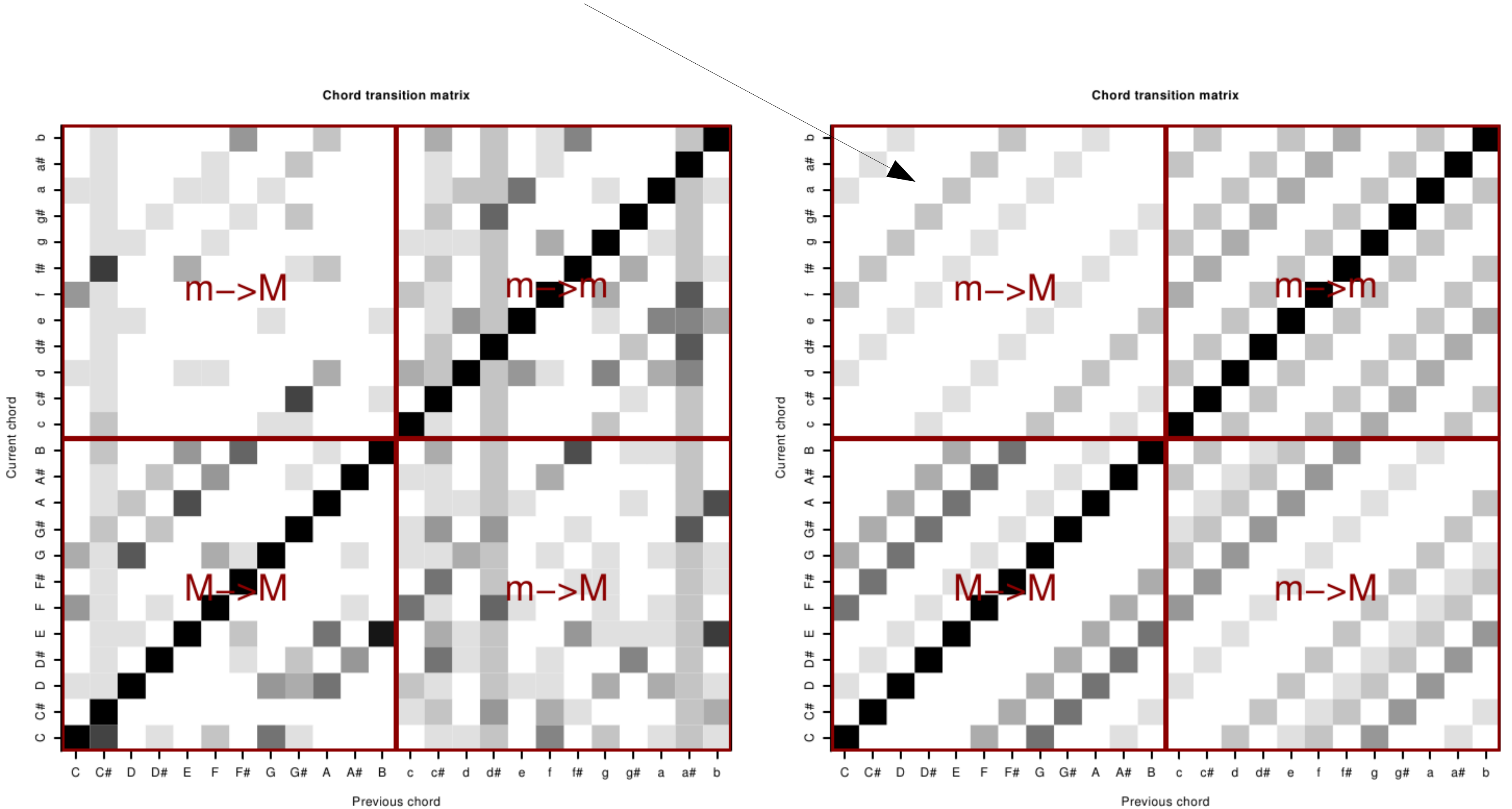


1-1 sequences are less likely than 0-1

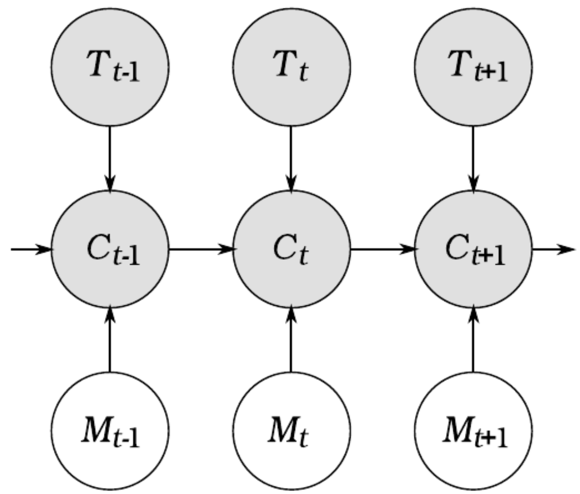
Chord model

$$P(C_t | C_{t-1})$$

- Modeled with a multinomial distribution.
- 24-chord dictionary.
- State-tying is used because we do not model the tonality.



Harmonization submodels



Melody $\mathrm{P}_1 = \mathrm{P}(C_t | M_t)$

Tonality $\mathrm{P}_2 = \mathrm{P}(C_t | T_t)$

Chord bigram $\mathrm{P}_3 = \mathrm{P}(C_t | C_{t-1})$

Note: it is a *discriminative* model

Melody submodel

$$\mathrm{P}_1 = \mathrm{P}(C_t | M_t)$$

- M_t is a set of active notes at time frame t.
- State tying: note patterns with the same content relative to the chord root were given identical probabilities, *e.g.*, the unordered note combination (C,G) in the chord of C-major is equally probable as the note combination (D♯,A♯) in the chord of D♯-major

Chord bigram submodel $\mathrm{P}_3 = \mathrm{P}(C_t | C_{t-1})$

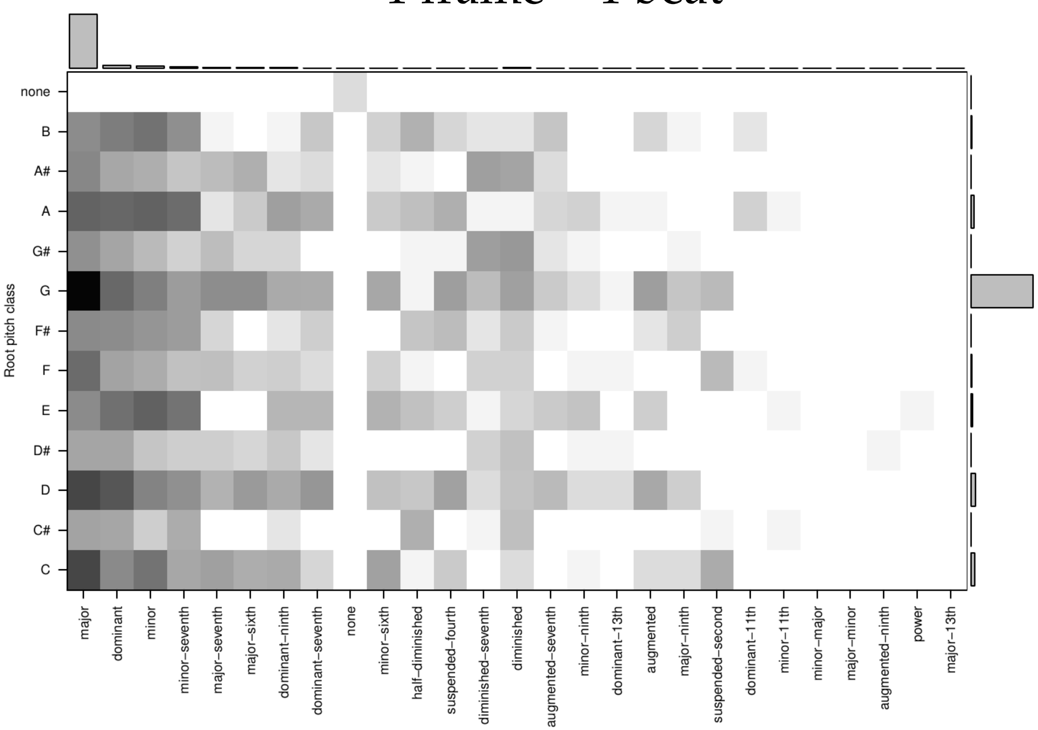
- A binary trigram model in the frequency domain.
- Chord labelled by one of 13 root pitch classes:
  C, C♯, D, D♯, E, F, F♯, G, G♯, A, A♯, B or "none" for non-chords

  and one of 27 chord types:

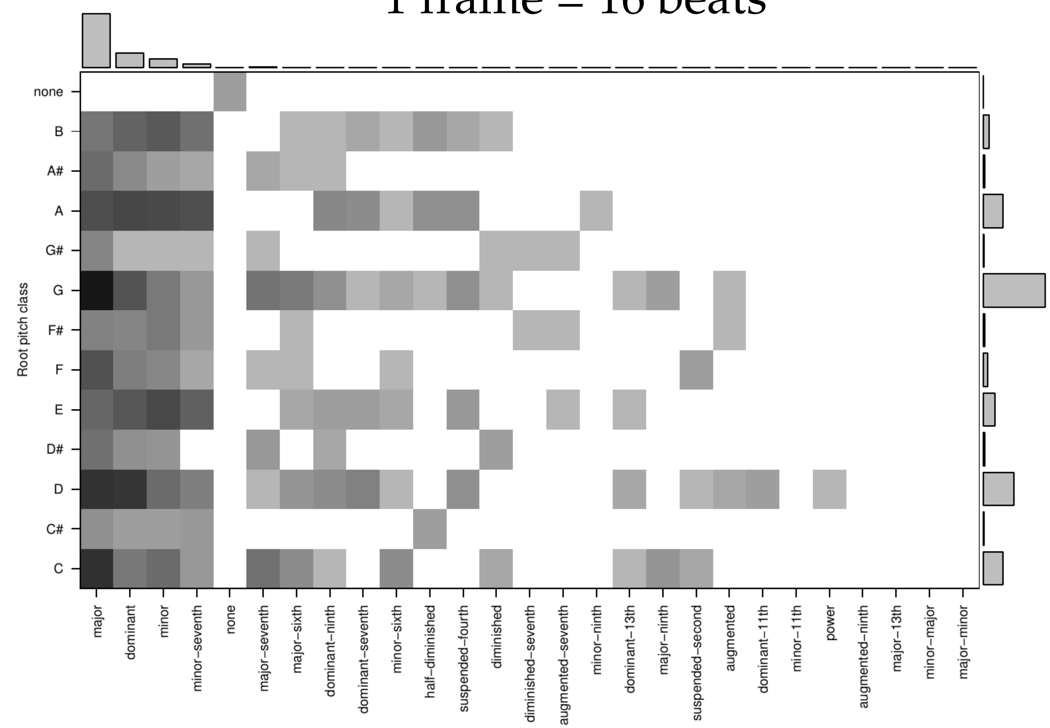
  major, minor, dominant, diminished, half-diminished, augmented, power, suspended-second, suspended-fourth, major-sixth, minor-sixth, major-seventh, minor-seventh, dominant-seventh, diminished-seventh, augmented-seventh, major-ninth, minor-ninth, dominant-ninth, augmented-ninth, minor-eleventh, dominant-eleventh, major-minor, minor-major, major-thirteenth, dominant-thirteenth or "none" for non-chords
- N = 351 distinct chord labels

Chord bigram submodel $\mathrm{P}_3 = \mathrm{P}(C_t | C_{t-1})$

1 frame = 1 beat



1 frame = 16 beats

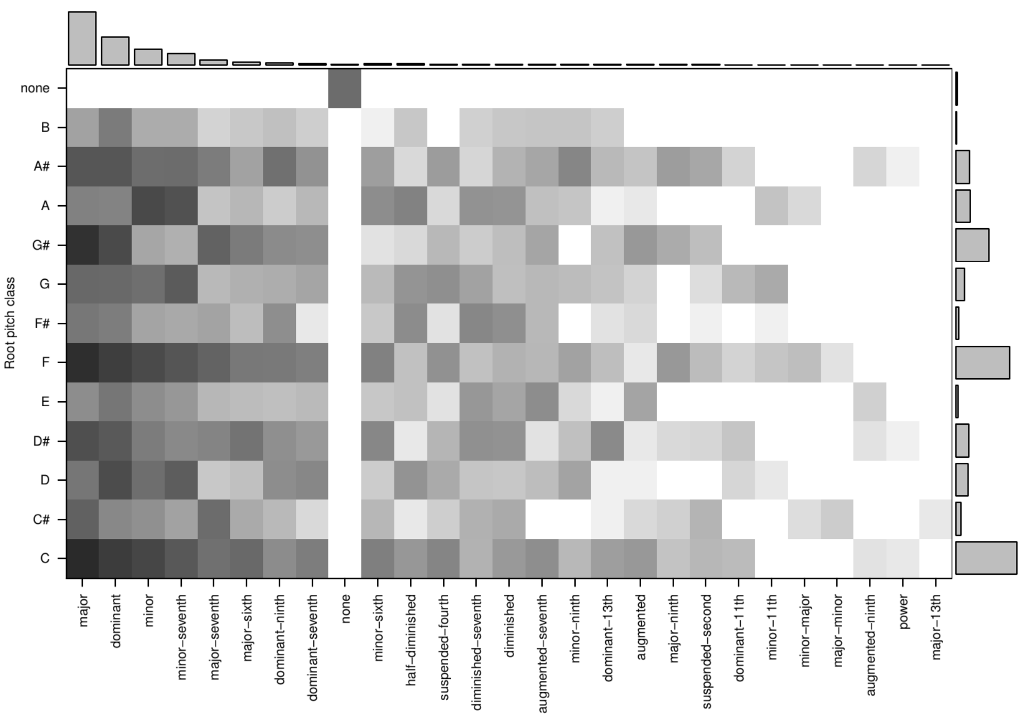


$C_{t\text{-}1}$ = G-maj

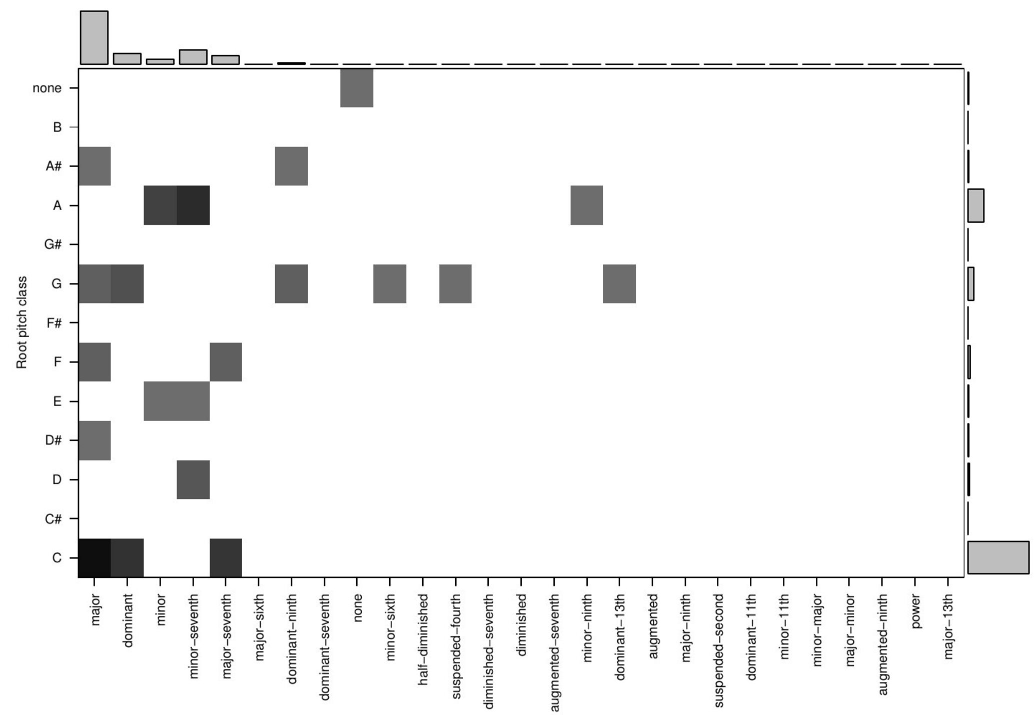
Melody submodel

$$\mathrm{P}_1 = \mathrm{P}(C_t | M_t)$$

$M_t = (\mathrm{C})$



$M_t = (\mathrm{C,E,G})$



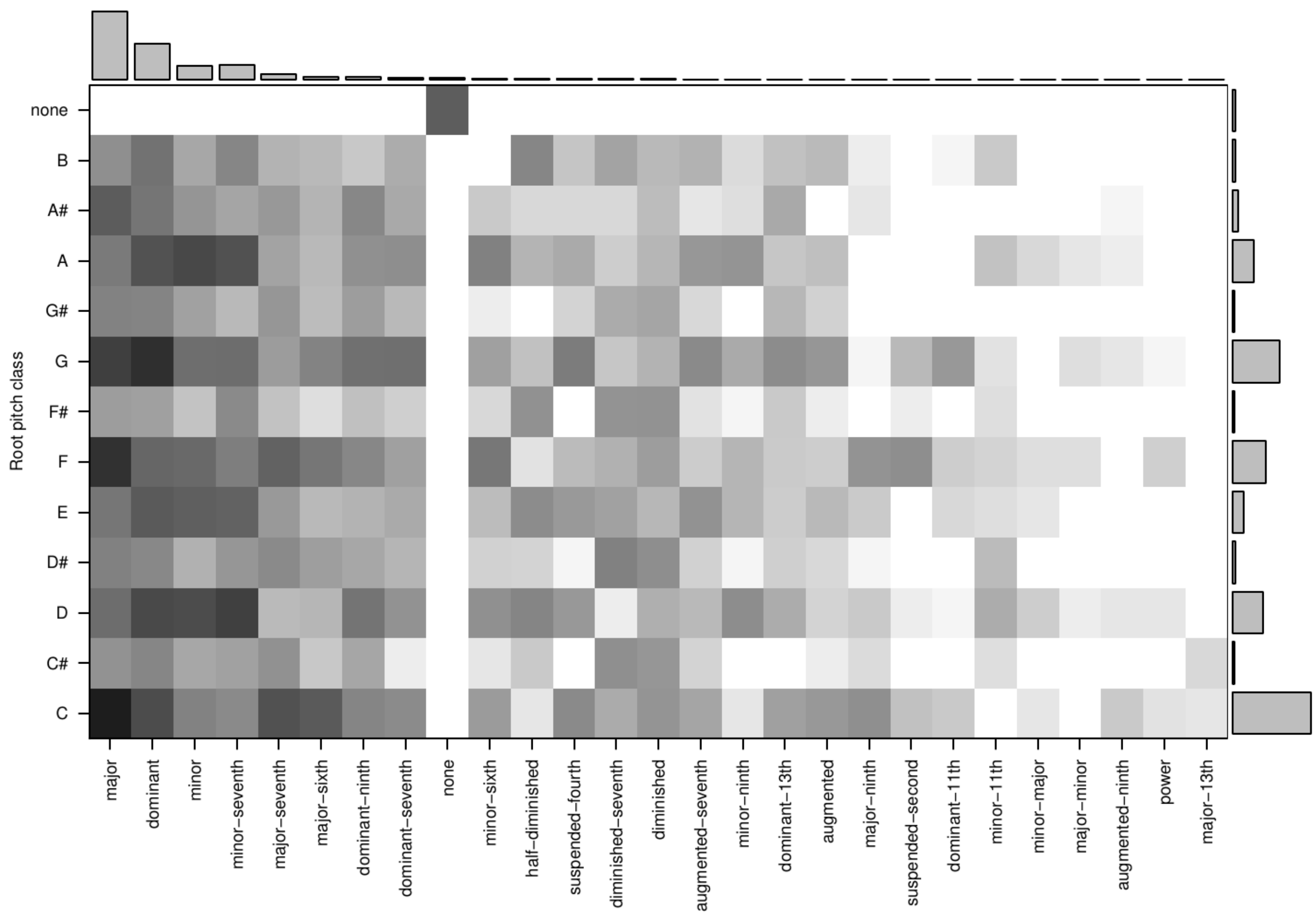
1 frame = 1 beat

Tonality submodel

$$\mathrm{P}_2 = \mathrm{P}(C_t|T_t)$$

- Tonality encoded as one of 24 different key labels resulting from the combination of 12 tonics (C, C♯, D, D♯, E, F, F♯, G, G♯, A, A♯, B) and 2 modes (major or minor)

- State tying: chords corresponding to the same scale degree in different keys are tied together.



1 frame = 1 beat

T_t = C-maj

Smoothing

- To avoid overfitting in the submodels, they are interpolated with simpler chord models (*additive smoothing*): chord *unigram* and *zero-gram*:

$$\mathrm{P}(C_t|\mathbf{C}_{1:t-1}, \mathbf{X}_{1:t}) = \alpha \mathrm{P}(C_t) + \beta + \sum_{i=1}^{I} a_i \mathrm{P}_i(C_t|\mathbf{A}_{i,t})$$

$$\mathbf{A}_{i,t} \subset \{C_{1:t-1}, \mathbf{X}_{1:t}\}$$

Subset of variables　　Full set of variables

$$\alpha + \beta + \sum_{i=1}^{I} a_i = 1$$

Smoothing

- In case of log-linear interpolation, each submodel is smoothed separately:

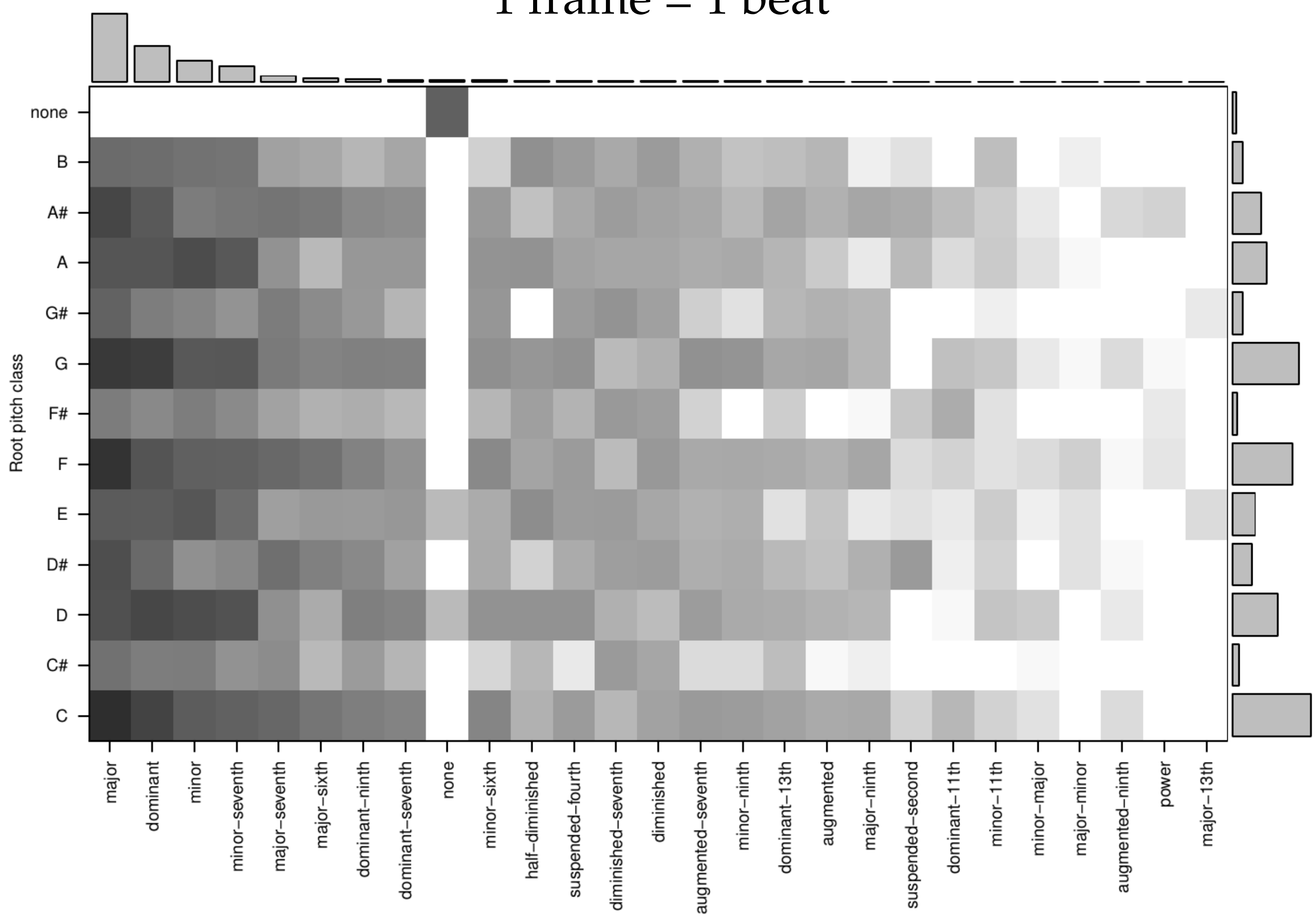
$$\mathrm{P}(C_t|\mathbf{C}_{1:t-1}, \mathbf{X}_{1:t}) = Z^{-1} \prod_{i=1}^{I} (\gamma_i \mathrm{P}_i(C_t|\mathbf{A}_{i,t}) + \delta_i \mathrm{P}(C_t) + \epsilon_i)^{b_i}$$

$$\gamma_i + \delta_i + \epsilon_i = 1 \qquad \text{for all } i$$

$$Z = \sum_{C_t} \prod_{i=1}^{I} (\gamma_i \mathrm{P}_i(C_t|\mathbf{A}_{i,t}) + \delta_i \mathrm{P}(C_t) + \epsilon_i)^{b_i}$$

Chord unigram submodel $\mathrm{P}(C_t)$

1 frame = 1 beat



Evaluation

Multiple pitch analysis data

- Mutopia dataset was used:
  - ~1300 files for training model parameters
  - 100 fles for validation
  - 100 files for testing
  - 1 frame = 1/6 of a beat
- RWC files annotated with harmony was used to train the harmony submodel and the chord models

Harmonization data

- For training, we have used a collection of around 2000 lead sheets from the Wikifonia web page:
  - melodies annotated with keys and absolute chord labels,
  - mostly popular (e.g., pop, rock) songs from the twentieth and the twenty-first centuries,
  - the songs were first screened for improper chord labels and wrong keys.

Training

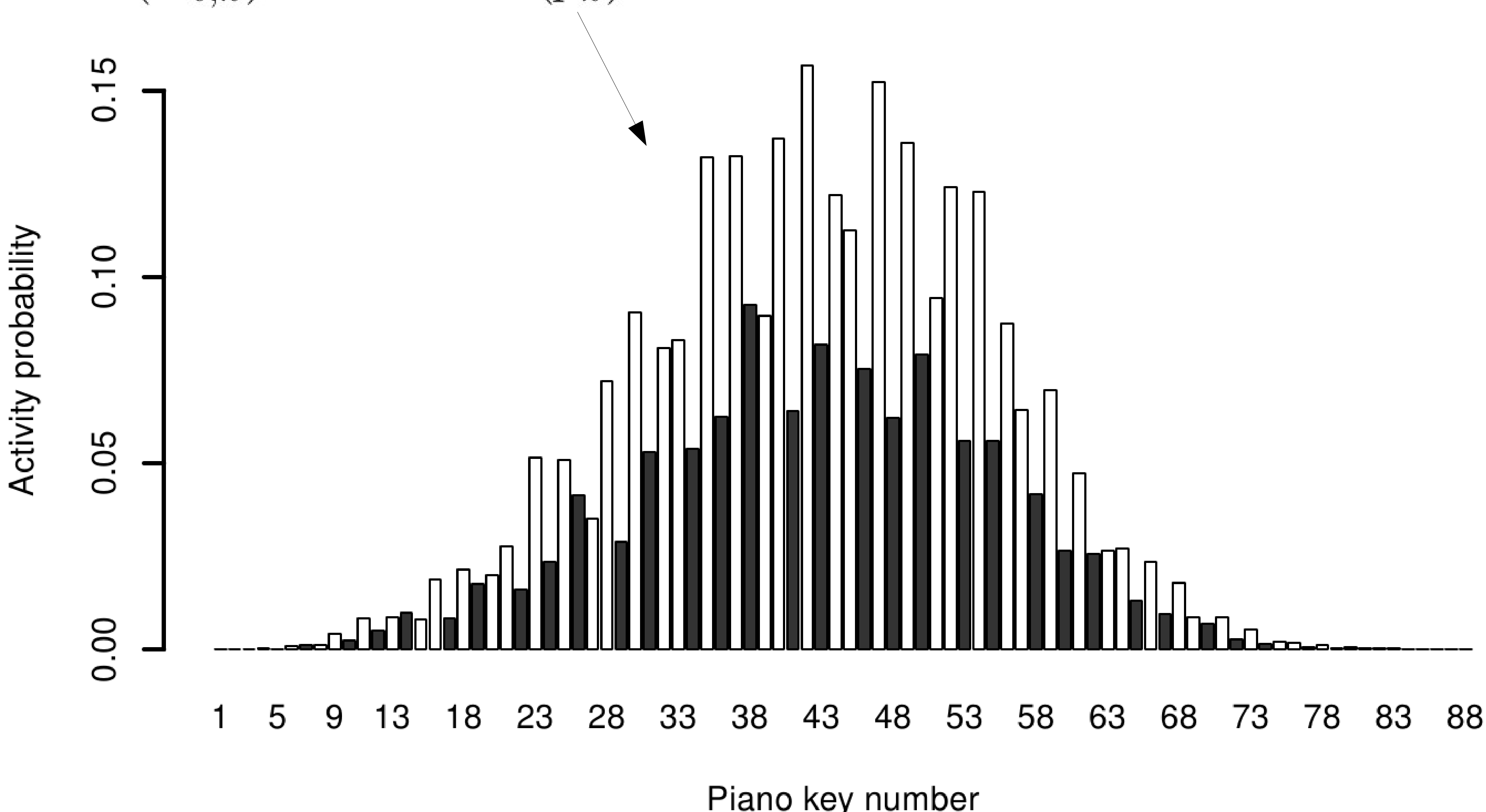
- Model parameters were trained by counting occurrences (maximizing the likelihood) on the *training dataset*.
- The smoothing parameters were optimized by maximizing the average cross-entropy of individual submodels on the *validation dataset*.
- Interpolation coefficients and smoothing for linear-combined harmonization model were optimized by maximizing cross-entropy of the *validation dataset*

$$\widehat{\lambda} = \arg\max_{\lambda} \log \mathrm{P}(\mathbf{N}|\lambda)$$

Reference pitch model

$\mathrm{P}(N_{t,k}) \sim \mathrm{Bernoulli}(p) \qquad p = 0.03807$ **B**

$\mathrm{P}(N_{t,k}) \sim \mathrm{Bernoulli}(p_k)$ **PB**



Cross-entropy

- Common metric for measuring modeling power of language [7] and music [5,6] models.
- Multipitch estimation:

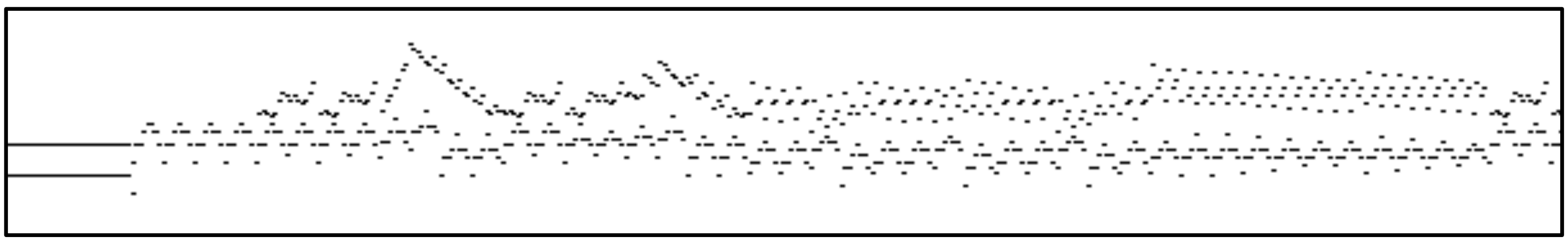
$$\begin{aligned} \mathrm{H}(\Lambda) &= -\frac{1}{KT}\log_2 \mathrm{P}(\mathbf{N}|\Lambda) \\ &= -\frac{1}{88T}\log_2 \sum_{\mathbf{C}} \mathrm{P}(\mathbf{N}|\mathbf{C},\Lambda)\mathrm{P}(\mathbf{C}|\Lambda) \end{aligned}$$

- Harmonization:

$$\mathrm{H}(\Lambda) = -\frac{1}{T}\log_2 \mathrm{P}(\mathbf{C}|\Lambda) = -\frac{1}{T}\log_2\left(\mathrm{P}(C_t|M_1,T_1)\prod_{t=2}^{T}\mathrm{P}(C_t|C_{t-1},M_t,T_t)\right)$$

Contextual cross-entropy

- For multipitch analysis, the cross-entropy value is dominated by the silence (97% notes are inactive on average).



- We would like to know how well do the models model the note activity, *i.e.*, note onsets, note offsets and notes – *contextual cross-entropy*.

$$\mathrm{cH}(\Lambda) = -\frac{1}{\sum_{t=1}^{T} |S_t|} \sum_{t=1}^{T} \sum_{k \in S_t} \log_2 \mathrm{P}(N_{t,k} | \mathbf{N}_{t-1}, N_{t,1:k-1})$$

Pitch cross-entropy

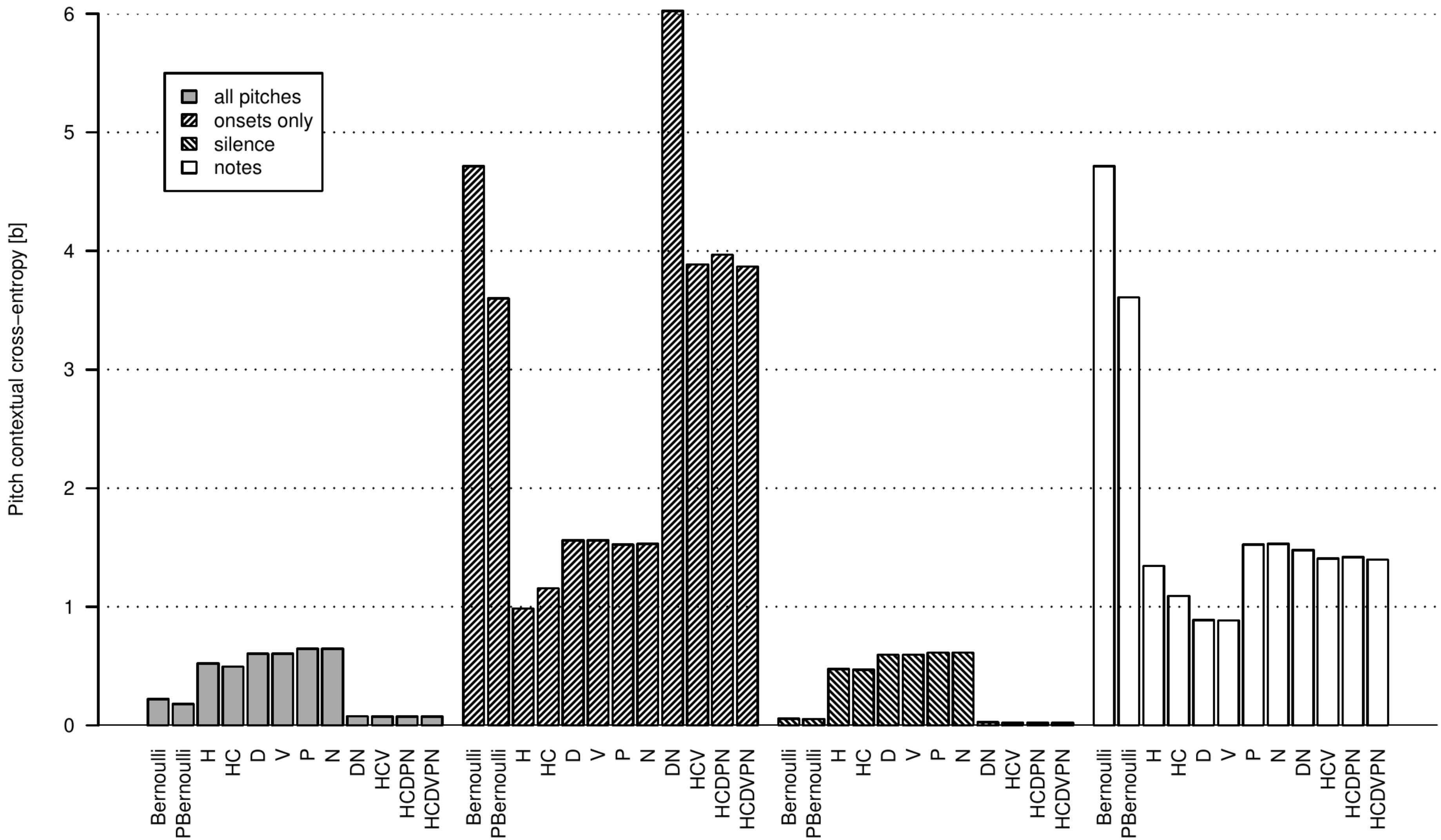
- Regular cross-entropy (in milibits):

	DN	HCV	HCDPN	HCDVPN
Linear	605.3	76.5	77.2	75.8
Log-linear	77.1	73.4	74.6	73.1
Difference	528.2	3.1	2.6	2.7

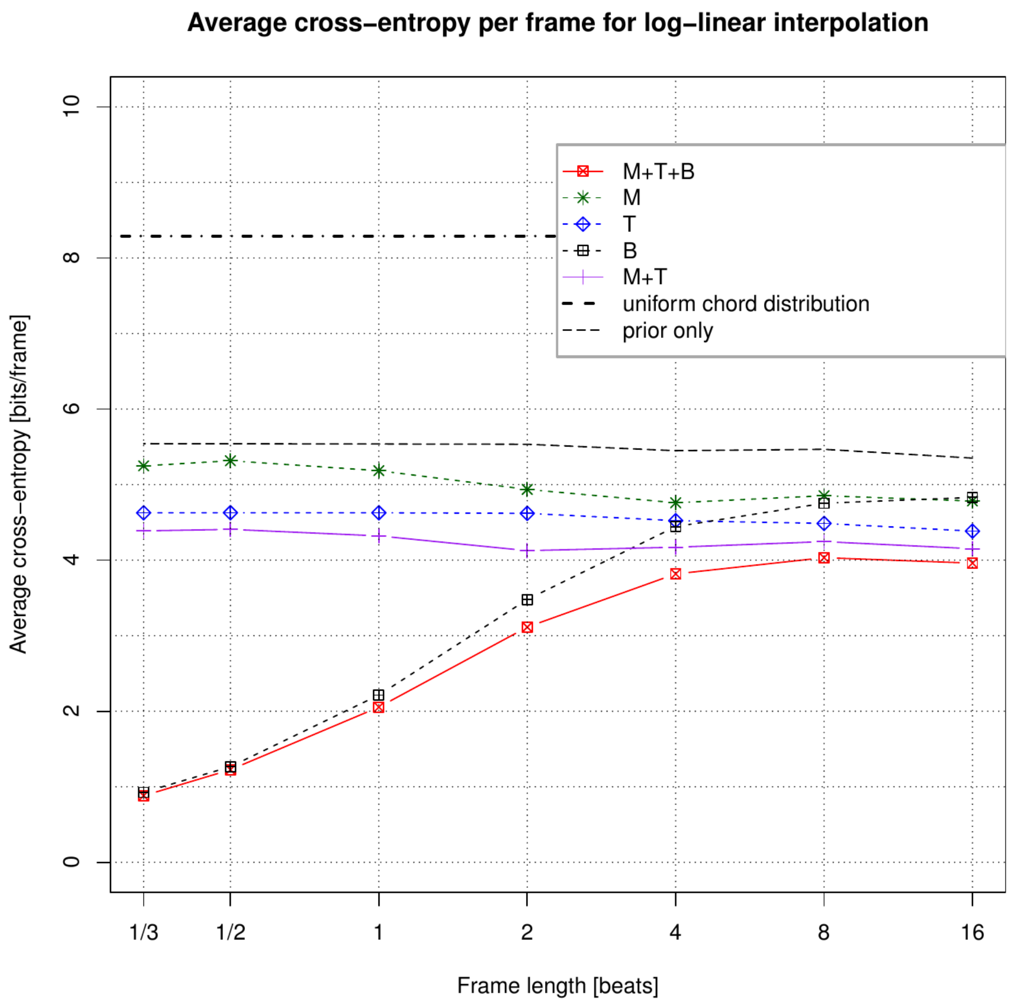
- Contextual cross-entropy (in milibits):

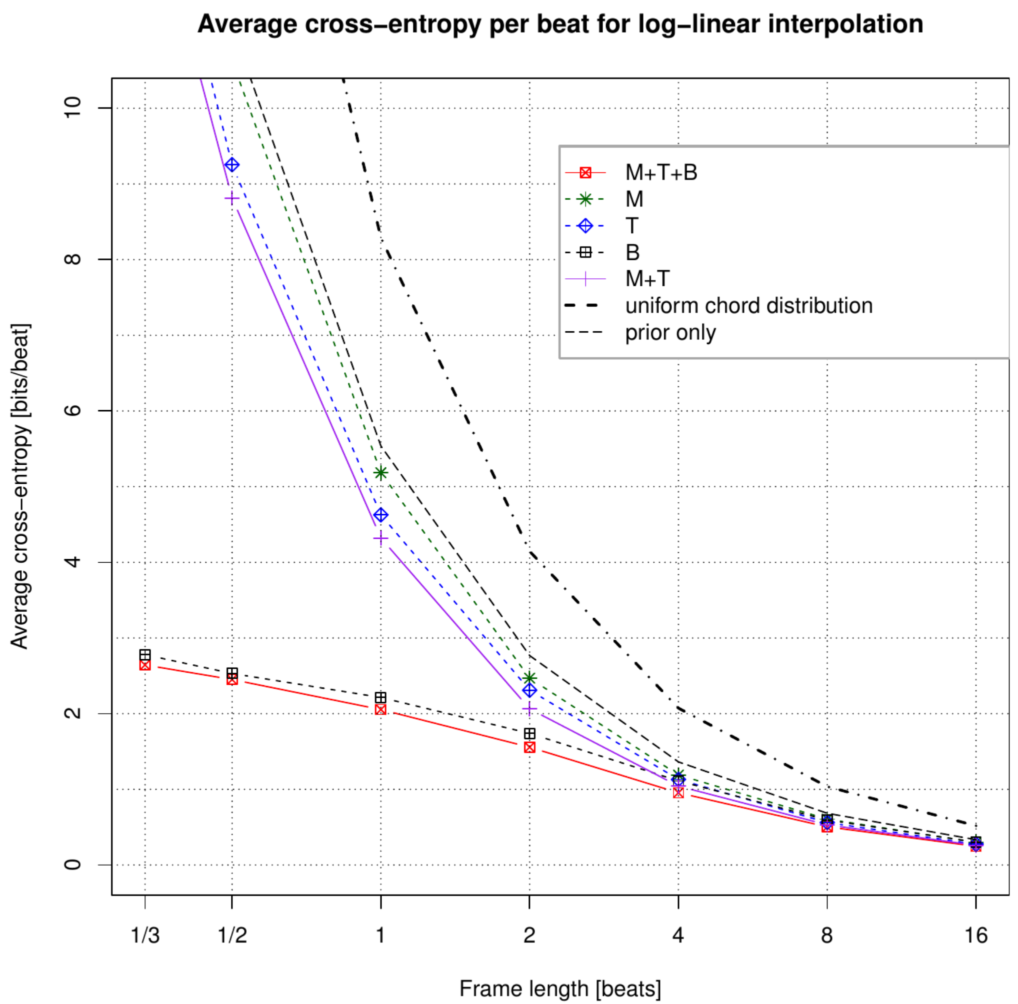
	DN	HCV	HCDPN	HCDVPN
Linear	1,560.0	4,042.7	4,058.9	3,963.4
Log-linear	6,022.7	3,886.3	3,969.5	3,869.7
Difference	-4462.7	156.4	89.4	93.7

Pitch cross-entropy



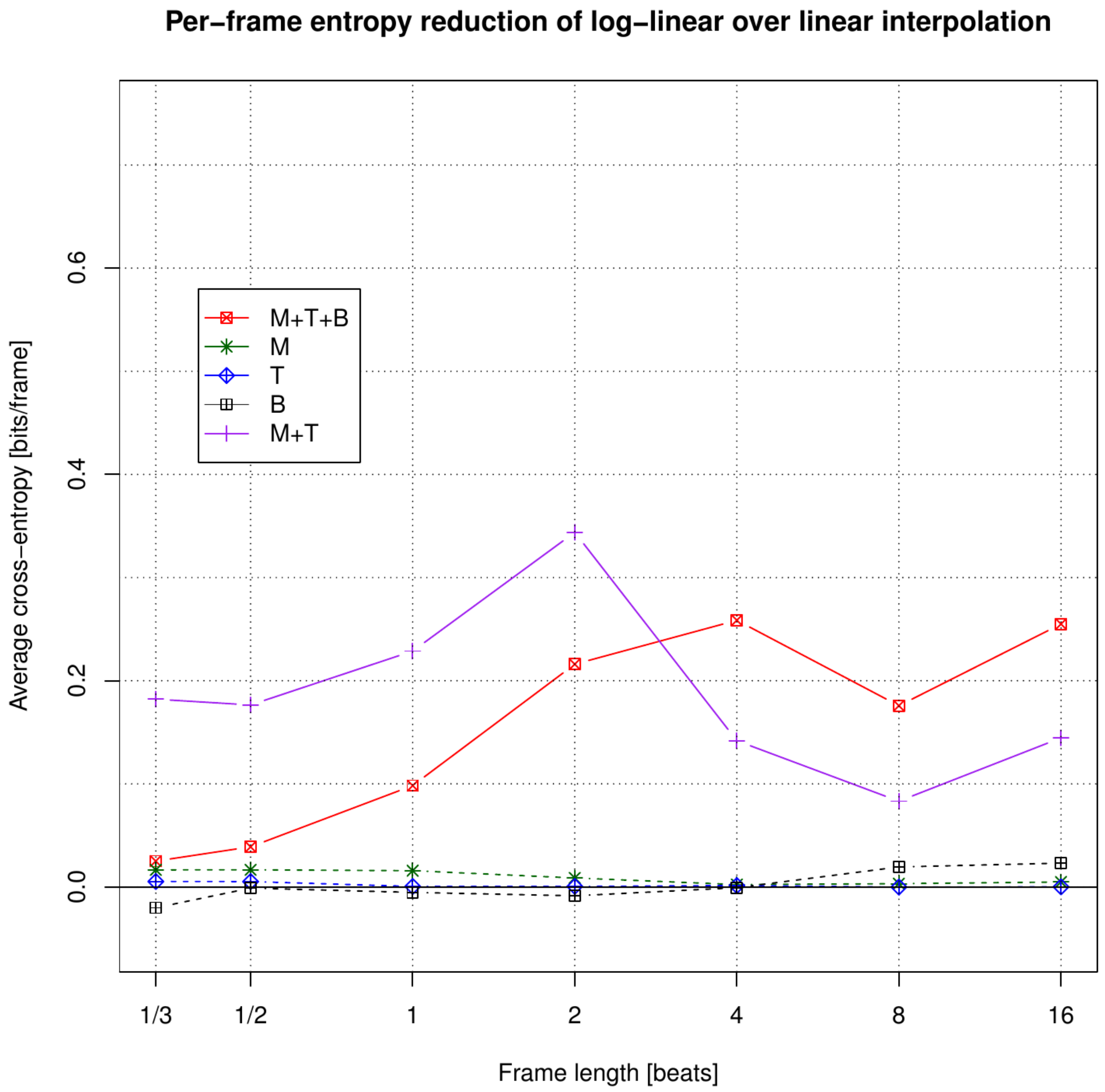
Harmonization cross-entropy





M = melody submodel, T = tonality submodel, B = chord bigram submodel

Harmonization cross-entropy



Accuracy

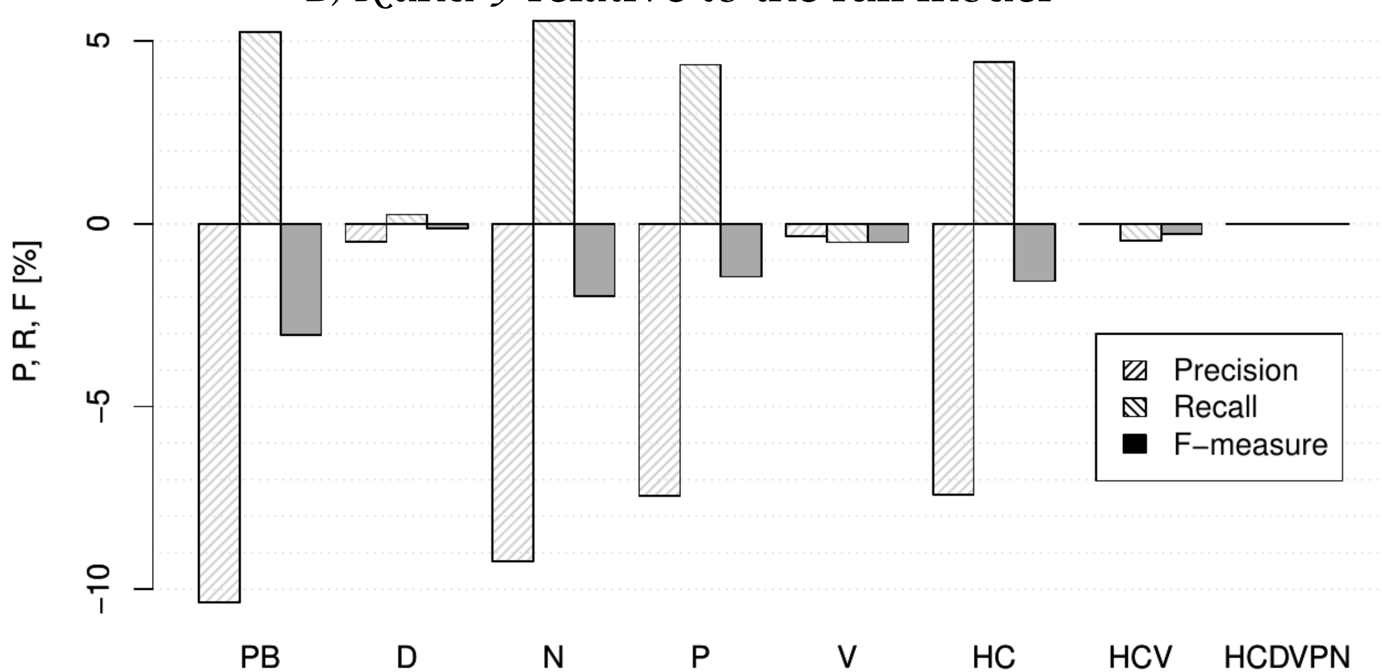
- Multipitch estimation:
  - Precision, Recall and F-measure
  - Reference musicological model: Bernoulli (equivalent to thresholded NMF) and pitch-dependent Bernoulli (eq. to pitch-dependent threshold)
- Harmonization:
  - Root note estimation accuracy (compared to leadsheets) and triad accuracy (root note + first chord interval)
  - Reference musicological model: Harmonic Analyzer by Temperley & Sleator [34]

Pitch estimation accuracy

Precision $\mathcal{P}$, Recall $\mathcal{R}$ and F-measure $\mathcal{F}$

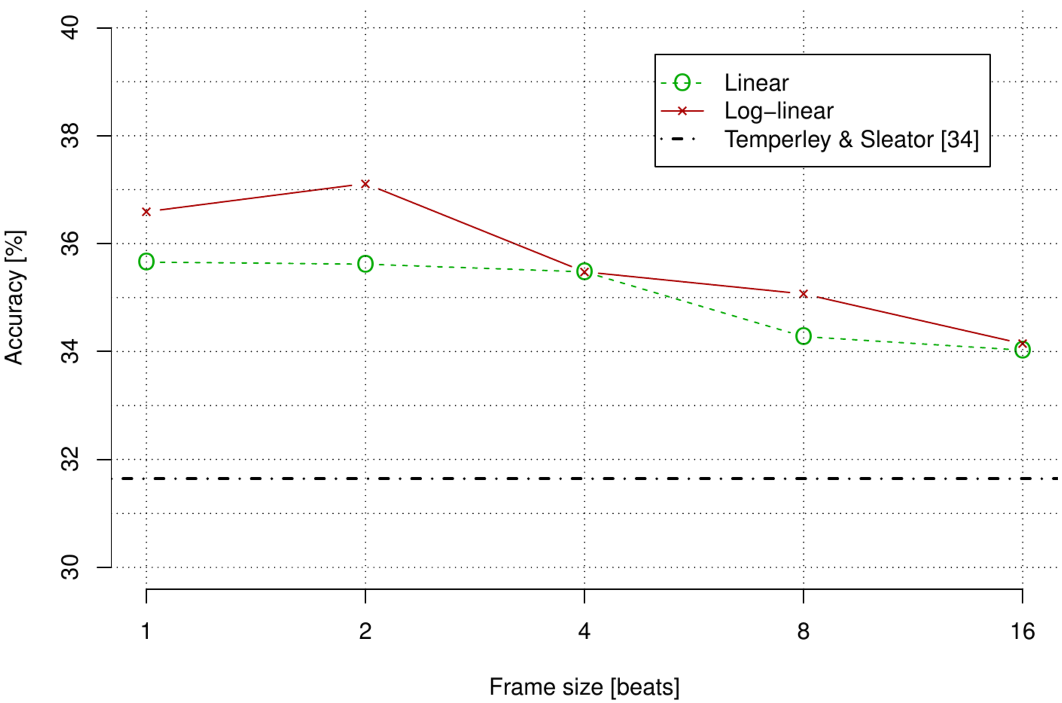
	PB	D	N	P	V	HC	HCV	HCDVPN
$\mathcal{P}$	73.0%	82.9%	74.2%	76.0%	83.1%	76.0%	83.4%	83.4%
$\mathcal{R}$	83.6%	78.7%	83.9%	82.7%	77.9%	82.8%	77.9%	78.4%
$\mathcal{F}$	76.1%	79.1%	77.2%	77.7%	78.7%	77.6%	78.9%	79.2%

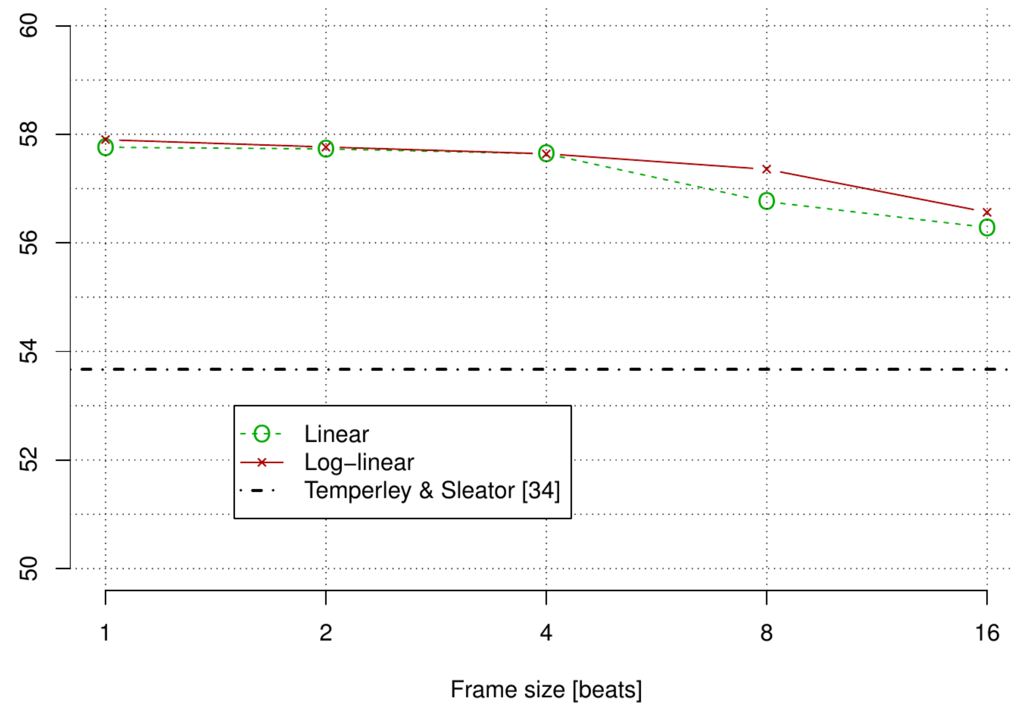
$\mathcal{P}$, $\mathcal{R}$ and $\mathcal{F}$ relative to the full model



Harmonization accuracy

- Root note estimation accuracies



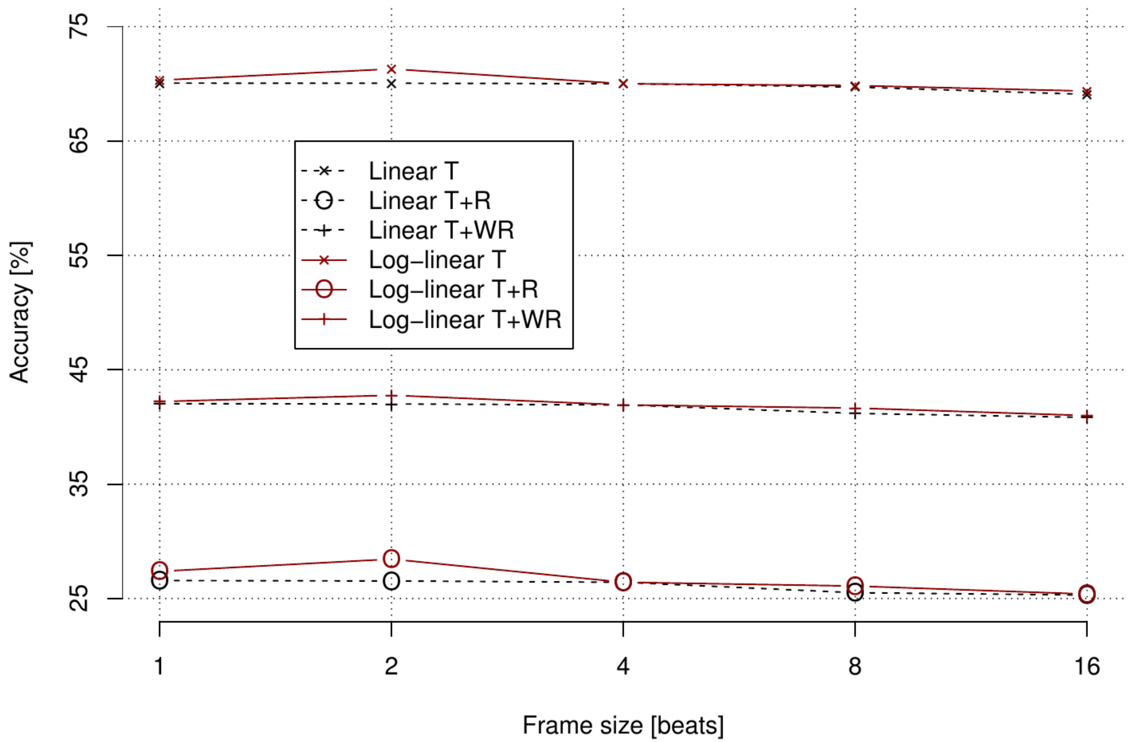


simple

weighter

Harmonization accuracy

- Triad accuracies



WR = weighted root note accuracy

Conclusion

- Multiple musical variables can be jointly modeled to improve their estimates
- Model interpolation is efficient in dealing with joint model complexity
- Linear interpolation seems to work slightly worse than the log-linear one

Possible future work

- A larger number of more complex sub-models could be investigated for further improvement in terms of cross-entropy and accuracy.
- Proposed method could be tested on a larger populations of songs that would include more diverse musical genres.
- Subjective listening tests could also be used to analyze the quality of the harmonizations in more detail.
- Model interpolation could be applied to other MIR tasks that would potentially benefit from modeling several musical aspects simultaneously.

Thank you!

References

[1] E. Vincent, S. Raczyński, N. Ono, and S. Sagayama, "*A roadmap towards versatile MIR*," in Proc. 11th International Conference on Music Information Retrieval (ISMIR), 2010, pp. 662–664.

[2] Simon, I., Morris, D., & Basu, S. (2008). "*MySong: automatic accompaniment generation for vocal melodies*." In Proc. 26th SIGCHI Conference on Human Factors in Computing Systems (pp. 725–734).

[3] PG Music Inc. (2012, August). "*Band-in-a-box*." http://www.pgmusic.com/.

[4] Klakow, D. (1998). "*Log-linear interpolation of language models*." In Proc. 5th International Conference on Spoken Language Processing (pp. 1695–1698).

[5] Allan, M., & Williams, C. (2005). "*Harmonising chorales by probabilistic inference*." Advances in Neural Information Processing Systems, 17 , 25–32.

[6] Paiement, J., Eck, D., & Bengio, S. (2006). "*Probabilistic melodic harmonization*." In Proc. 19th Canadian Conf. on Artificial Intelligence (pp. 218–229).

[7] Kneser, R.; Ney, H., "*Improved backing-off for M-gram language modeling*," Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on , vol.1, no., pp.181,184 vol.1, 9-12 May 1995

[34] Temperley, D., & Sleator, D. (2012, August). "*Harmonic Analyzer*." http://www.cs.cmu.edu/~sleator/harmonic-analysis/.