

# Cognitive Music Modelling: An Information Dynamics Approach

**Samer Abdallah, Henrik Ekeus, Peter Foster,  
Andrew Robertson and Mark Plumbley**

Centre for Digital Music  
Queen Mary, University of London

June 1, 2012

# Outline

**Expectation and surprise in music**

**Surprise, entropy and information in random sequences**

**Markov chains**

**Application: The Melody Triangle**

**More process models**

**Application: Analysis of minimalist music**

**Application: Beat tracking and rhythm**

**Summary and conclusions**

# Outline

## **Expectation and surprise in music**

Surprise, entropy and information in random sequences

Markov chains

Application: The Melody Triangle

More process models

Application: Analysis of minimalist music

Application: Beat tracking and rhythm

Summary and conclusions

# 'Unfoldingness'

Music is experienced as a

# 'Unfoldingness'

Music is experienced as a phenomenon

# ‘Unfoldingness’

Music is experienced as a phenomenon that

# ‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’

# ‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’ in



# 'Unfoldingness'

Music is experienced as a phenomenon that 'unfolds' in  
blancmange

# 'Unfoldingness'

Music is experienced as a phenomenon that 'unfolds' in  
(just kidding)

# ‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’ in time,

# ‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’ in time, rather than being apprehended as a static object presented in its entirety.

# ‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’ in time, rather than being apprehended as a static object presented in its entirety.

[This is recognised in computation linguistics where the phenomenon is known as *incrementality*, e.g. in incremental parsing.]

# ‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’ in time, rather than being apprehended as a static object presented in its entirety.

[This is recognised in computation linguistics where the phenomenon is known as *incrementality*, e.g. in incremental parsing.]

Meyer [Mey67] argued that musical experience depends on how we change and revise our conceptions *as events happen*, on how expectation and prediction interact with occurrence, and that, to a large degree, the way to understand the effect of music is to focus on this ‘kinetics’ of expectation and surprise.

# Expectation and surprise in music

Music creates *expectations* of what is to come next, which may be fulfilled immediately, after some delay, or not at all. Suggested by music theorists, e.g. L. B. Meyer [Mey67] and Narmour [Nar77] but also noted much earlier by Hanslick [Han86] in the 1850s:

*'The most important factor in the mental process which accompanies the act of listening to music, and which converts it to a source of pleasure, is ... the intellectual satisfaction which the listener derives from continually following and anticipating the composer's intentions—now, to see his expectations fulfilled, and now, to find himself agreeably mistaken. It is a matter of course that this intellectual flux and reflux, this perpetual giving and receiving takes place unconsciously, and with the rapidity of lightning-flashes.'*

# Probabilistic reasoning

Making predictions and assessing surprise is essentially reasoning with degrees of belief and (arguably) the best way to do this is using Bayesian probability theory [Cox46, Jay88].

[NB. this is **subjective** probability as advocated by e.g. De Finetti and Jaynes.]



# Probabilistic reasoning

Making predictions and assessing surprise is essentially reasoning with degrees of belief and (arguably) the best way to do this is using Bayesian probability theory [Cox46, Jay88].

[NB. this is **subjective** probability as advocated by e.g. De Finetti and Jaynes.]

We suppose that familiarity with different styles of music takes the form of various probabilistic models, and that these models are adapted through listening.

# Probabilistic reasoning

Making predictions and assessing surprise is essentially reasoning with degrees of belief and (arguably) the best way to do this is using Bayesian probability theory [Cox46, Jay88].

[NB. this is **subjective** probability as advocated by e.g. De Finetti and Jaynes.]

We suppose that familiarity with different styles of music takes the form of various probabilistic models, and that these models are adapted through listening.

Experimental evidence that humans are able to internalise statistical knowledge about musical: [SJAN99, ETK02]; and also that statistical models are effective for computational analysis of music, e.g. [CW95, Pea05].

# Music and information theory

With probabilistic models in hand we can apply quantitative information theory: we can compute entropies, relative entropies, mutual information, and all that.

# Music and information theory

With probabilistic models in hand we can apply quantitative information theory: we can compute entropies, relative entropies, mutual information, and all that.

Lots of interest in application of information theory to perception, music and aesthetics since the 50s, e.g. Moles [Mol66], Meyer [Mey67], Cohen [Coh62], Berlyne [Ber71]. (See also Bense, Hiller)

# Music and information theory

With probabilistic models in hand we can apply quantitative information theory: we can compute entropies, relative entropies, mutual information, and all that.

Lots of interest in application of information theory to perception, music and aesthetics since the 50s, e.g. Moles [Mol66], Meyer [Mey67], Cohen [Coh62], Berlyne [Ber71]. (See also Bense, Hiller)

Idea is that subjective qualities and states like uncertainty, surprise, complexity, tension, and interestingness are determined by information-theoretic quantities.

# Music and information theory

With probabilistic models in hand we can apply quantitative information theory: we can compute entropies, relative entropies, mutual information, and all that.

Lots of interest in application of information theory to perception, music and aesthetics since the 50s, e.g. Moles [Mol66], Meyer [Mey67], Cohen [Coh62], Berlyne [Ber71]. (See also Bense, Hiller)

Idea is that subjective qualities and states like uncertainty, surprise, complexity, tension, and interestingness are determined by information-theoretic quantities.

Berlyne [Ber71] called such quantities ‘collative variables’, since they are to do with patterns of occurrence rather than medium-specific details. *Information aesthetics*.

# Probabilistic model-based observer hypothesis

- As we listen, we maintain a probabilistic model that enables us to make predictions. As events unfold, we revise our probabilistic 'belief state', including predictions about the future.

# Probabilistic model-based observer hypothesis

- As we listen, we maintain a probabilistic model that enables us to make predictions. As events unfold, we revise our probabilistic 'belief state', including predictions about the future.
- Probability distributions and changes in distributions are characterised in terms of information theoretic-measures such as entropy and relative entropy (KL divergence).



# Probabilistic model-based observer hypothesis

- As we listen, we maintain a probabilistic model that enables us to make predictions. As events unfold, we revise our probabilistic ‘belief state’, including predictions about the future.
- Probability distributions and changes in distributions are characterised in terms of information theoretic-measures such as entropy and relative entropy (KL divergence).
- The dynamic evolution of these information measures captures significant structure, e.g. events that are surprising, informative, explanatory etc.

# Features of information dynamics

**Abstraction:** sensitive mainly to *patterns* of occurrence, rather than details of which specific things occur or the sensory medium.

# Features of information dynamics

**Abstraction:** sensitive mainly to *patterns* of occurrence, rather than details of which specific things occur or the sensory medium.

**Generality:** applicable in principle to any probabilistic model, in particular, models with time-dependent latent variables such as HMMs. Many important musical concepts like key, harmony, and beat are essentially ‘hidden variables’.

# Features of information dynamics

**Abstraction:** sensitive mainly to *patterns* of occurrence, rather than details of which specific things occur or the sensory medium.

**Generality:** applicable in principle to any probabilistic model, in particular, models with time-dependent latent variables such as HMMs. Many important musical concepts like key, harmony, and beat are essentially ‘hidden variables’.

**Richness:** when applied to models with latent variables, can result in many-layered analysis, capturing information flow about harmony, tempo, etc.

# Features of information dynamics

**Abstraction:** sensitive mainly to *patterns* of occurrence, rather than details of which specific things occur or the sensory medium.

**Generality:** applicable in principle to any probabilistic model, in particular, models with time-dependent latent variables such as HMMs. Many important musical concepts like key, harmony, and beat are essentially ‘hidden variables’.

**Richness:** when applied to models with latent variables, can result in many-layered analysis, capturing information flow about harmony, tempo, etc.

**Subjectivity:** all probabilities are *subjective* probabilities relative to *observer's* model, which can depend on observer's capabilities and prior experience.

# Outline

Expectation and surprise in music

**Surprise, entropy and information in random sequences**

Markov chains

Application: The Melody Triangle

More process models

Application: Analysis of minimalist music

Application: Beat tracking and rhythm

Summary and conclusions

# Information theory primer · Entropy

Let  $X$  be a discrete-valued random (in the sense of *subjective* probability) variable. Entropy is a measure of *uncertainty*. If observer expects to see  $x$  with probability  $p(x)$ , then

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} -p(x) \log p(x) \\ &= \mathbb{E}[-\log p(X)]. \end{aligned}$$

Consider  $-\log p(x)$  as the ‘surprisingness’ of  $x$ , then the entropy is the ‘expected surprisingness’. High for spread out distributions and low for concentrated ones.

# Information theory primer · Relative entropy

Relative entropy or Kullback-Leibler (KL) divergence quantifies difference between probability distributions. If observer receives data  $\mathcal{D}$ , divergence between (subjective) prior and posterior distributions is the amount of information in  $\mathcal{D}$  *about*  $X$  for this observer:

$$I(\mathcal{D} \rightarrow X) = D(p_{X|\mathcal{D}} || p_X) = \sum_{x \in \mathcal{X}} p(x|\mathcal{D}) \log \frac{p(x|\mathcal{D})}{p(x)}.$$

If observing  $\mathcal{D}$  causes a large change in belief about  $X$ , then  $\mathcal{D}$  contained a lot of information about  $X$ .

Like Lindley's (1956) information (thanks Lars!).



# Information theory primer · Mutual information

Mutual information between (MI)  $X_1$  and  $X_2$  is the expected amount of information about  $X_2$  in an observation of  $X_1$ . Can be written in several ways:

$$\begin{aligned} I(X_1; X_2) &= \sum_{x_1, x_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} \\ &= H(X_1) + H(X_2) - H(X_1, X_2) \\ &= H(X_2) - H(X_2|X_1). \end{aligned}$$

- (1) Expected information about  $X_2$  in an observation of  $X_1$ ;
- (2) Expected reduction in uncertainty about  $X_2$  after observing  $X_1$ ;
- (3) Symmetric:  $I(X_1; X_2) = I(X_2; X_1)$ .

## Information theory primer · Conditional MI

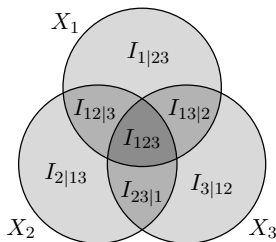
Information in one variable about another given observations of some third variable. Formulated analogously by adding conditioning variables to entropies:

$$I(X_1; X_2 | X_3) = H(X_1 | X_3) - H(X_1 | X_2, X_3).$$

Makes explicit the dependence of information assessment on background knowledge, represented by conditioning variables.

# Information theory primer · I-Diagrams

Information diagrams are a Venn diagram-like representation of entropies and mutual informations for a set of random variables.



$$I_{1|23} = H(X_1|X_2, X_3)$$

$$I_{13|2} = I(X_1; X_3|X_2)$$

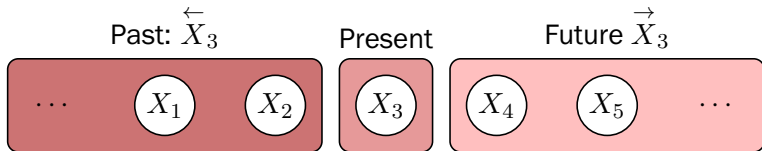
$$I_{1|23} + I_{13|2} = H(X_1|X_2)$$

$$I_{12|3} + I_{123} = I(X_1; X_2)$$

The areas of the three circles represent  $H(X_1)$ ,  $H(X_2)$  and  $H(X_3)$  respectively. The total shaded area is the joint entropy  $H(X_1, X_2, X_3)$ . Each undivided region is an *atom* of the I-diagram.

# Information theory in sequences

Consider an observer receiving elements of a random sequence  $(\dots, X_{-1}, X_0, X_1, X_2, \dots)$ , so that at any time  $t$  there is a 'present'  $X_t$ , an observed past  $\overleftarrow{X}_t$ , and an unobserved future  $\overrightarrow{X}_t$ . Eg, at time  $t = 3$ :



Consider how the observer's belief state evolves when, having observed up to  $X_2$ , it learns the value of  $X_3$ .

# 'Surprise' based quantities

To obtain first set of measures, we ignore the future  $\vec{X}_t$  and consider the probability distribution for  $X_t$  give the observed past  $\overleftarrow{X}_t = \overleftarrow{x}_t$ .

**1 Surprisingness:** negative log-probability  $\ell_t = -\log p(x_t | \overleftarrow{x}_t)$ .

# 'Surprise' based quantities

To obtain first set of measures, we ignore the future  $\vec{X}_t$  and consider the probability distribution for  $X_t$  give the observed past  $\overleftarrow{X}_t = \overleftarrow{x}_t$ .

- 1 **Surprisingness:** negative log-probability  $\ell_t = -\log p(x_t | \overleftarrow{x}_t)$ .
- 2 Expected surprisingness given context  $\overleftarrow{X} = \overleftarrow{x}_t$  is the entropy of the predictive distribution,  $H(X_t | \overleftarrow{X}_t = \overleftarrow{x}_t)$ : uncertainty about  $X_t$  before the observation is made.

# 'Surprise' based quantities

To obtain first set of measures, we ignore the future  $\vec{X}_t$  and consider the probability distribution for  $X_t$  given the observed past  $\overleftarrow{X}_t = \overleftarrow{x}_t$ .

- 1 **Surprisingness:** negative log-probability  $\ell_t = -\log p(x_t | \overleftarrow{x}_t)$ .
- 2 Expected surprisingness given context  $\overleftarrow{X} = \overleftarrow{x}_t$  is the entropy of the predictive distribution,  $H(X_t | \overleftarrow{X}_t = \overleftarrow{x}_t)$ : uncertainty about  $X_t$  before the observation is made.
- 3 Expectation over all possible realisations of process is the conditional entropy  $H(X_t | \overleftarrow{X}_t)$  according to the observer's model. For stationary process, is *entropy rate*  $h_\mu$ .

# Predictive information

Second set of measures based on amount of information the observation  $X_t = x_t$  carries *about* about the unobserved future  $\vec{X}_t$ , given that we already know the past  $\overleftarrow{X}_t = \overleftarrow{x}_t$ : is

$$\mathcal{I}_t = I(X_t = x_t \rightarrow \vec{X}_t | \overleftarrow{X}_t = \overleftarrow{x}_t).$$

Is KL divergence between beliefs about future  $\vec{X}_t$  prior and posterior to observation  $X_t = x_t$ . Hence, for continuous valued variables, invariant to invertible transformations of the observation spaces.



# Predictive information based quantities

- 1 *Instantaneous predictive information (IPI) is just  $\mathcal{I}_t$ .*

# Predictive information based quantities

- 1 *Instantaneous predictive information* (IPI) is just  $\mathcal{I}_t$ .
- 2 Expectation of  $\mathcal{I}_t$  before observation at time  $t$  is  $I(X_t; \vec{X}_t | \overleftarrow{X}_t = \overleftarrow{x}_t)$ : mutual information conditioned on observed past. Is the amount of new information about the future expected from the next observation. Useful for directing attention towards the next event even before it happens?

# Predictive information based quantities

- 1 *Instantaneous predictive information* (IPI) is just  $\mathcal{I}_t$ .
- 2 Expectation of  $\mathcal{I}_t$  before observation at time  $t$  is  $I(X_t; \vec{X}_t | \overleftarrow{X}_t = \overleftarrow{x}_t)$ : mutual information conditioned on observed past. Is the amount of new information about the future expected from the next observation. Useful for directing attention towards the next event even before it happens?
- 3 Expectation over all possible realisations is the conditional mutual information  $I(X_t; \vec{X}_t | \overleftarrow{X}_t)$ . For stationary process, this is the *global predictive information rate* (PIR), the average rate at which new information arrives about the future. In terms of conditional entropies, has two forms:  
 $H(\vec{X}_t | \overleftarrow{X}_t) - H(\vec{X}_t | X_t, \overleftarrow{X}_t)$  or  $H(X_t | \overleftarrow{X}_t) - H(X_t | \vec{X}_t, \overleftarrow{X}_t)$ .

# Global measures for stationary processes

For a stationary random process model, the average levels of surprise and information are captured by the time-shift invariant process information measures:

$$\text{entropy rate : } h_\mu = H(X_t | \overleftarrow{X}_t)$$

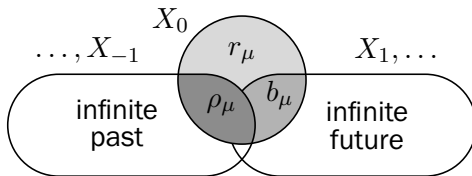
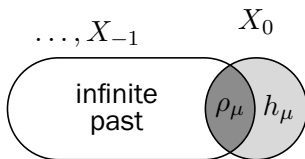
$$\text{multi-information rate : } \rho_\mu = I(\overleftarrow{X}_t; X_t) = H(X_t) - h_\mu$$

$$\text{residual entropy rate : } r_\mu = H(X_t | \overleftarrow{X}_t, \overrightarrow{X}_t)$$

$$\text{predictive information rate : } b_\mu = I(X_t; \overrightarrow{X}_t | \overleftarrow{X}_t) = h_\mu - r_\mu$$

Residual entropy also known as *erasure entropy* [VW06].

# Process I-diagrams



Marginal entropy of 'present'  $X_0$  is  $H(X_0) = \rho_\mu + r_\mu + b_\mu$ .  
 Entropy rate is  $h_\mu = r_\mu + b_\mu$ .

# Outline

Expectation and surprise in music

Surprise, entropy and information in random sequences

**Markov chains**

Application: The Melody Triangle

More process models

Application: Analysis of minimalist music

Application: Beat tracking and rhythm

Summary and conclusions

# Markov chains · Definitions

Let  $X$  be a Markov chain with state space  $\{1, \dots, K\}$ , i.e. the  $X_t$  take values from 1 to  $K$ .



Parameterised by transition matrix  $a \in \mathbb{R}^{K \times K}$ , i.e.

$p(X_{t+1}=i|X_t=j) = a_{ij}$ . Assume irreducibility, ergodicity etc. to ensure uniqueness of stationary distribution  $\pi$  such that

$p(X_t=i) = \pi_i^a$  independent of  $t$ . Entropy rate as a function of  $a$  is

$$h(a) = \sum_{j=1}^K \pi_j^a \sum_{i=1}^K -a_{ij} \log a_{ij}.$$

# Markov chains · PIR

Predictive information rate for first order chains comes out in terms of entropy rate function as

$$b_{\mu} = h(a^2) - h(a),$$

where  $a^2$  is two-step transition matrix.



# Markov chains · PIR

Predictive information rate for first order chains comes out in terms of entropy rate function as

$$b_\mu = h(a^2) - h(a),$$

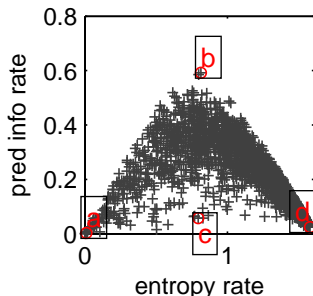
where  $a^2$  is two-step transition matrix.

Can be generalised to higher-order transition matrices

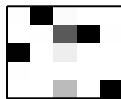
$$b_\mu = h(\hat{a}^{N+1}) - Nh(\hat{a}),$$

where  $N$  is the order of the chain and  $\hat{a}$  is a sparse  $K^N \times K^N$  transition matrix over product state space of  $N$  consecutive observations (step size 1).

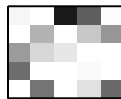
# Entropy rate and PIR in Markov chains



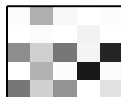
transmat (a)



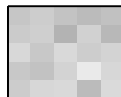
transmat (b)



transmat (c)



transmat (d)



For given  $K$ , entropy rate varies between 0 (deterministic sequence) and  $\log K$  when  $a_{ij} = 1/K$  for all  $i, j$ . Space of transition matrices explored by generating them at random and plotting entropy rate vs PIR. (Note inverted 'U' relationship).

# Samples from processes with different PIR

sequence (a)



sequence (b)



sequence (c)



sequence (d)



Sequence (a) is repetition of state 4 (see transmat (a) on previous slide). System (b) has the highest PIR.

# Outline

Expectation and surprise in music

Surprise, entropy and information in random sequences

Markov chains

**Application: The Melody Triangle**

More process models

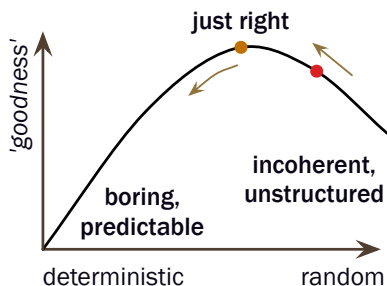
Application: Analysis of minimalist music

Application: Beat tracking and rhythm

Summary and conclusions

# Complexity and interestingness: the Wundt Curve

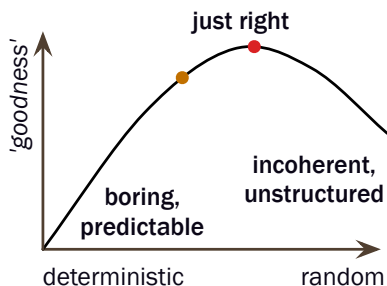
Studies looking into the relationship between stochastic complexity (usually measured as entropy or entropy rate) and aesthetic value, reveal an inverted 'U' shaped curve [Ber71]. (Also, Wundt curve [Wun97]). Repeated exposure tends to move stimuli leftwards.



Explanations for this usually appeal to a need for a 'balance' between order and chaos, unity and diversity, and so on, in a generally imprecise way.

# Complexity and interestingness: the Wundt Curve

Studies looking into the relationship between stochastic complexity (usually measured as entropy or entropy rate) and aesthetic value, reveal an inverted 'U' shaped curve [Ber71]. (Also, Wundt curve [Wun97]). Repeated exposure tends to move stimuli leftwards.



Explanations for this usually appeal to a need for a 'balance' between order and chaos, unity and diversity, and so on, in a generally imprecise way.

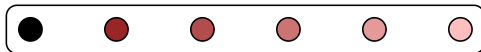
# PIR as a measure of cognitive activity

The predictive information rate incorporates a similar balance automatically: is maximal for sequences which are neither deterministic nor totally uncorrelated across time.

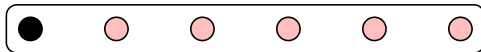
too predictable:



intermediate:



too random:



(Black: *observed*; red: *unobserved*; paler: *greater uncertainty*.) Our interpretation: Things are 'interesting' or at least 'salient' when each new part supplies new information about parts to come.

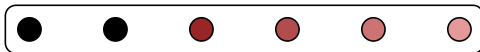
# PIR as a measure of cognitive activity

The predictive information rate incorporates a similar balance automatically: is maximal for sequences which are neither deterministic nor totally uncorrelated across time.

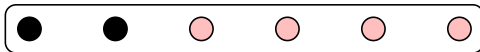
too predictable:



intermediate:



too random:



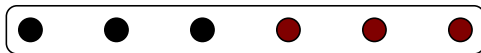
(Black: *observed*; red: *unobserved*; paler: *greater uncertainty*.) Our interpretation: Things are 'interesting' or at least 'salient' when each new part supplies new information about parts to come.



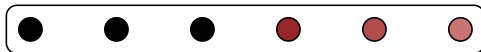
# PIR as a measure of cognitive activity

The predictive information rate incorporates a similar balance automatically: is maximal for sequences which are neither deterministic nor totally uncorrelated across time.

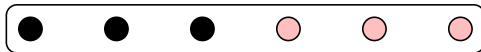
too predictable:



intermediate:



too random:

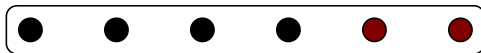


(Black: *observed*; red: *unobserved*; paler: *greater uncertainty*.) Our interpretation: Things are ‘interesting’ or at least ‘salient’ when each new part supplies new information about parts to come.

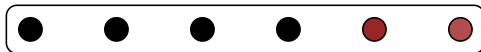
# PIR as a measure of cognitive activity

The predictive information rate incorporates a similar balance automatically: is maximal for sequences which are neither deterministic nor totally uncorrelated across time.

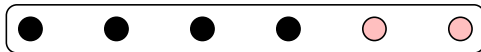
too predictable:



intermediate:

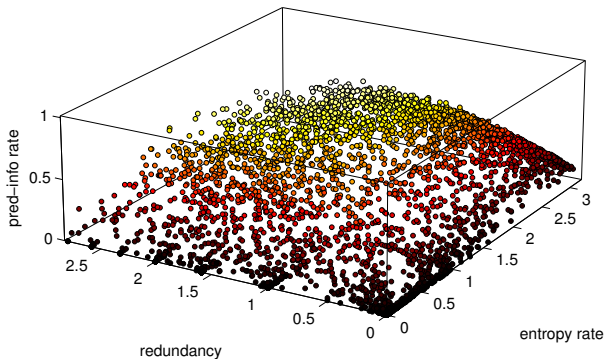


too random:



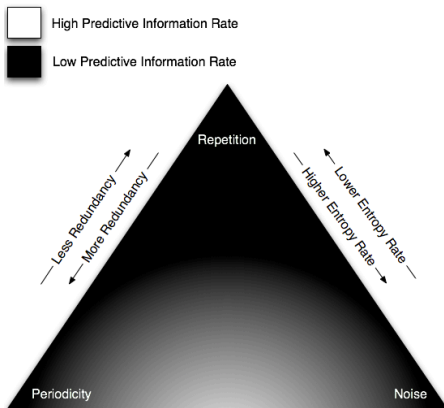
(Black: *observed*; red: *unobserved*; paler: *greater uncertainty*.) Our interpretation: Things are ‘interesting’ or at least ‘salient’ when each new part supplies new information about parts to come.

# The Melody Triangle · Information space



Population of transition matrices in 3D space of  $h_\mu$ ,  $\rho_\mu$  and  $b_\mu$ .  
Colour of each point represents PIR. Shape is mostly (not completely) hollow inside: forming roughly a curved triangular sheet.

# The Melody Triangle · User interface



Allows user to place tokens in the triangle to cause sonification of a Markov chain with corresponding information ‘coordinate’.

## Subjective information

So far we've assumed that sequence is actually sampled from from a stationary Markov chain with a transition matrix known to the observer. This means time averages of IPI and surprise should equal expectations.

## Subjective information

So far we've assumed that sequence is actually sampled from from a stationary Markov chain with a transition matrix known to the observer. This means time averages of IPI and surprise should equal expectations.

What if sequence is sampled from some other Markov chain, or is produced by some unknown process?

# Subjective information

So far we've assumed that sequence is actually sampled from from a stationary Markov chain with a transition matrix known to the observer. This means time averages of IPI and surprise should equal expectations.

What if sequence is sampled from some other Markov chain, or is produced by some unknown process?

- In general, it may be impossible to identify any 'true' model. There are no 'objective' probabilities; only subjective ones, as argued by de Finetti [dF75].

# Subjective information

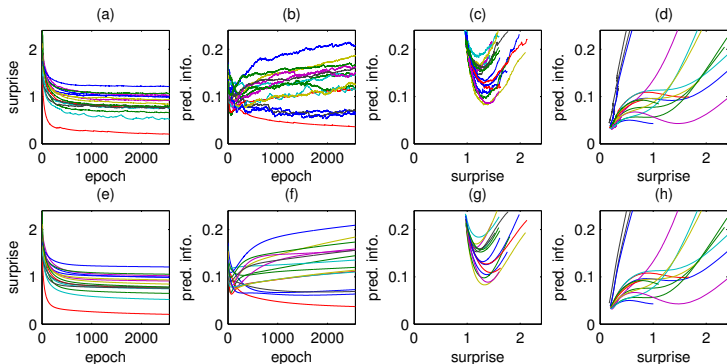
So far we've assumed that sequence is actually sampled from from a stationary Markov chain with a transition matrix known to the observer. This means time averages of IPI and surprise should equal expectations.

What if sequence is sampled from some other Markov chain, or is produced by some unknown process?

- In general, it may be impossible to identify any 'true' model. There are no 'objective' probabilities; only subjective ones, as argued by de Finetti [dF75].
- If sequence *is* sampled from some Markov chain, we can compute (time) averages of observer's average subjective surprise and PI and also track what happens if observer gradually learns the transition matrix from the data.



# Effect of learning on information dynamics



**(a/b/e/f):** multiple runs starting from same initial condition but using different generative transition matrices. **(c/d/g/h):** multiple runs starting from different initial conditions and converging on transition matrices with (c/g) high and (d/h) low PIR.

# Outline

Expectation and surprise in music

Surprise, entropy and information in random sequences

Markov chains

Application: The Melody Triangle

**More process models**

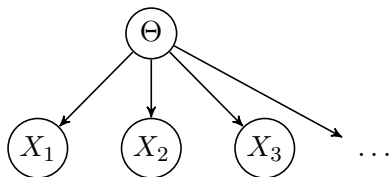
Application: Analysis of minimalist music

Application: Beat tracking and rhythm

Summary and conclusions

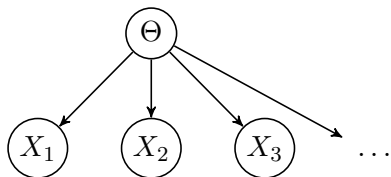
# Exchangeable sequences and parametric models

De Finetti's theorem says that an exchangeable random process can be represented as a sequence variables which are iid *given* some hidden probability distribution, which we can think of as a parameterised model:



# Exchangeable sequences and parametric models

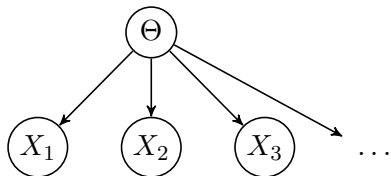
De Finetti's theorem says that an exchangeable random process can be represented as a sequence variables which are iid *given* some hidden probability distribution, which we can think of as a parameterised model:



Observer's belief state at time  $t$  includes probability distribution over the parameters  $p(\Theta = \theta | \overleftarrow{X}_t = \overleftarrow{x}_t)$ .

# Exchangeable sequences and parametric models

De Finetti's theorem says that an exchangeable random process can be represented as a sequence variables which are iid *given* some hidden probability distribution, which we can think of as a parameterised model:

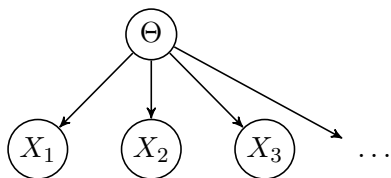


Observer's belief state at time  $t$  includes probability distribution over the parameters  $p(\Theta = \theta | \overleftarrow{X}_t = \overleftarrow{x}_t)$ .

Each observation causes revision of belief state and hence supplies information  $I(X_t = x_t \rightarrow \Theta | \overleftarrow{X}_t = \overleftarrow{x}_t)$  about  $\Theta$ : In previous work we called this the 'model information rate'.

# Exchangeable sequences and parametric models

De Finetti's theorem says that an exchangeable random process can be represented as a sequence variables which are iid *given* some hidden probability distribution, which we can think of as a parameterised model:

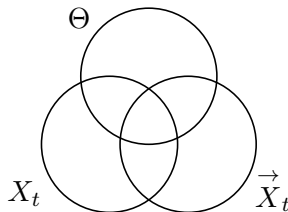


Observer's belief state at time  $t$  includes probability distribution over the parameters  $p(\Theta = \theta | \overleftarrow{X}_t = \overleftarrow{x}_t)$ .

Each observation causes revision of belief state and hence supplies information  $I(X_t = x_t \rightarrow \Theta | \overleftarrow{X}_t = \overleftarrow{x}_t)$  about  $\Theta$ : In previous work we called this the 'model information rate'. (Same as Haussler and Oppen's [H095] IIG or Itti and Baldi's [IB05] Bayesian surprise.)

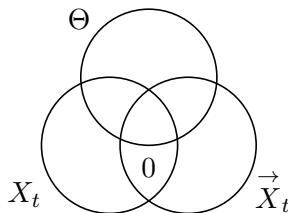
# IIG equals IPI in (some) XRP

Mild assumptions yield a relationship between IIG (instantaneous information gain) and IPI. (Everything here implicitly conditioned on  $\overleftarrow{X}_t$ ).



# IIG equals IPI in (some) XRP

Mild assumptions yield a relationship between IIG (instantaneous information gain) and IPI. (Everything here implicitly conditioned on  $\vec{X}_t$ ).

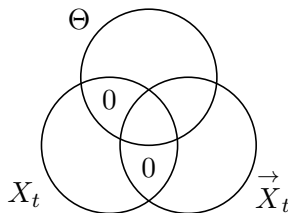


- 1  $X_t \perp \vec{X}_t | \Theta$ : observations iid given  $\Theta$  for XRP;



# IIG equals IPI in (some) XRP

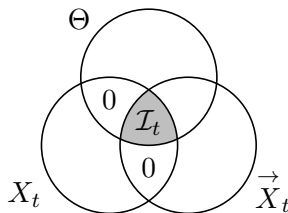
Mild assumptions yield a relationship between IIG (instantaneous information gain) and IPI. (Everything here implicitly conditioned on  $\vec{X}_t$ ).



- 1  $X_t \perp \vec{X}_t | \Theta$ : observations iid given  $\Theta$  for XRP;
- 2  $\Theta \perp X_t | \vec{X}_t$ : assumption that  $X_t$  adds no new information about  $\Theta$  given infinitely long sequence  $\vec{X}_t = X_{t+1:\infty}$ .

# IIG equals IPI in (some) XRP

Mild assumptions yield a relationship between IIG (instantaneous information gain) and IPI. (Everything here implicitly conditioned on  $\overleftarrow{X}_t$ ).

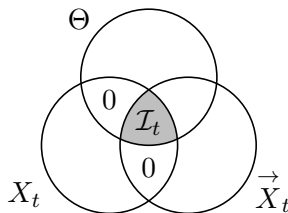


- 1  $X_t \perp \overrightarrow{X}_t | \Theta$ : observations iid given  $\Theta$  for XRP;
- 2  $\Theta \perp X_t | \overrightarrow{X}_t$ : assumption that  $X_t$  adds no new information about  $\Theta$  given infinitely long sequence  $\overrightarrow{X}_t = X_{t+1:\infty}$ .

Hence,  $I(X_t; \Theta_t | \overleftarrow{X}_t) = I(X_t; \overrightarrow{X}_t | \overleftarrow{X}_t) = \mathcal{I}_t$ .

# IIG equals IPI in (some) XRP

Mild assumptions yield a relationship between IIG (instantaneous information gain) and IPI. (Everything here implicitly conditioned on  $\overleftarrow{X}_t$ ).



- 1  $X_t \perp \overrightarrow{X}_t | \Theta$ : observations iid given  $\Theta$  for XRP;
- 2  $\Theta \perp X_t | \overrightarrow{X}_t$ : assumption that  $X_t$  adds no new information about  $\Theta$  given infinitely long sequence  $\overrightarrow{X}_t = X_{t+1:\infty}$ .

Hence,  $I(X_t; \Theta_t | \overleftarrow{X}_t) = I(X_t; \overrightarrow{X}_t | \overleftarrow{X}_t) = \mathcal{I}_t$ .

Can drop assumption 1 and still get  $I(X_t; \Theta_t | \overleftarrow{X}_t)$  as an additive component (lower bound) of  $\mathcal{I}_t$ .

# Discrete-time Gaussian processes

Information-theoretic quantities used earlier have analogues for continuous-valued random variables. For stationary Gaussian processes, we can obtain results in terms of the power spectral density  $S(\omega)$ , (which for discrete time is periodic in  $\omega$  with period  $2\pi$ ). Standard methods give

$$\begin{aligned} H(X_t) &= \frac{1}{2} \left( \log 2\pi e + \log \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) \, d\omega \right), \\ h_{\mu} &= \frac{1}{2} \left( \log 2\pi e + \frac{1}{2\pi} \int_{-\pi}^{\pi} \log S(\omega) \, d\omega \right), \\ \rho_{\mu} &= \frac{1}{2} \left( \log \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) \, d\omega - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log S(\omega) \, d\omega \right). \end{aligned}$$

Entropy rate is also known as Kolmogorov-Sinai entropy.

# PIR/Multi-information duality

Analysis yields PIR:

$$b_{\mu} = \frac{1}{2} \left( \log \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{S(\omega)} d\omega - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \frac{1}{S(\omega)} d\omega \right).$$

Yields simple expression for finite-order autoregressive processes,  
but beware: can diverge for moving average processes!

# PIR/Multi-information duality

Analysis yields PIR:

$$b_{\mu} = \frac{1}{2} \left( \log \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{S(\omega)} d\omega - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \frac{1}{S(\omega)} d\omega \right).$$

Yields simple expression for finite-order autoregressive processes, but beware: can diverge for moving average processes!

Compare with multi-information rate:

$$\rho_{\mu} = \frac{1}{2} \left( \log \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) d\omega - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log S(\omega) d\omega \right).$$

Yields simple expression for finite-order moving-average processes, but can diverge for marginally stable autoregressive processes.

# PIR/Multi-information duality

Analysis yields PIR:

$$b_{\mu} = \frac{1}{2} \left( \log \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{S(\omega)} d\omega - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \frac{1}{S(\omega)} d\omega \right).$$

Yields simple expression for finite-order autoregressive processes, but beware: can diverge for moving average processes!

Compare with multi-information rate:

$$\rho_{\mu} = \frac{1}{2} \left( \log \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) d\omega - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log S(\omega) d\omega \right).$$

Yields simple expression for finite-order moving-average processes, but can diverge for marginally stable autoregressive processes.

Infinities are troublesome and point to problem with notion of infinitely precise observation of continuous-valued variables.

# Outline

Expectation and surprise in music

Surprise, entropy and information in random sequences

Markov chains

Application: The Melody Triangle

More process models

**Application: Analysis of minimalist music**

Application: Beat tracking and rhythm

Summary and conclusions



# Material and Methods

We took two pieces of minimalist music by Philip Glass, *Two Pages* (1969) and *Gradus* (1968). Both monophonic and isochronous, so representable very simply as a sequence of symbols (notes), one symbol per beat, yet remain ecologically valid examples of ‘real’ music.

We use an elaboration of the Markov chain model—not necessarily a good model *per se*, but that wasn’t the point of the experiment. Markov chain model was chosen as it is tractable from an information dynamics point of view while not being completely trivial.

# Time-varying transition matrix model

We allow transition matrix to vary slowly with time to track changes in the sequence structure. Hence, observer's belief state includes a probability distribution over transition matrices; we choose a product of Dirichlet distributions:

$$p(a|\theta) = \prod_{j=1}^K p_{\text{Dir}}(a_{:j}|\theta_{:j}),$$

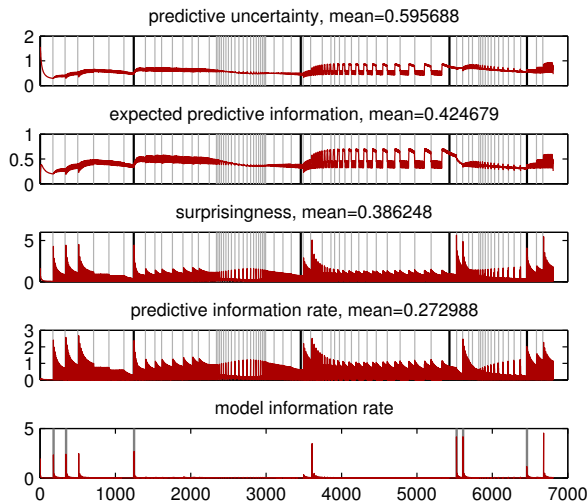
where  $a_{:j}$  is  $j^{\text{th}}$  column of  $a$  and  $\theta$  is an  $K \times K$  parameter matrix.

At each time step, distribution first *spreads* under mapping

$$\theta_{ij} \mapsto \frac{\beta \theta_{ij}}{(\beta + \theta_{ij})}$$

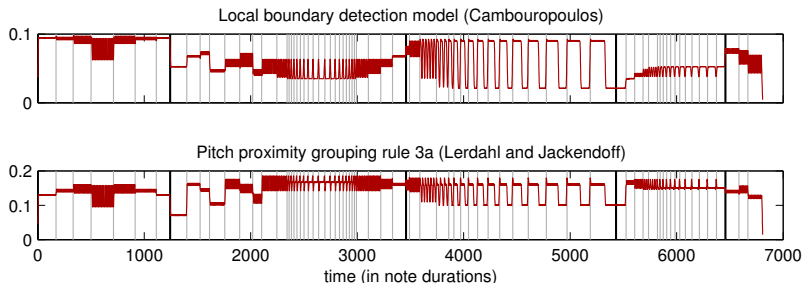
to model possibility that transition matrix has changed ( $\beta = 2500$  in our experiments). Then it *contracts* due to new observation providing fresh evidence about transition matrix.

# Two Pages · Results



**Thick lines:** part boundaries as indicated by Glass;  
**grey lines (top four panels):** changes in the melodic 'figures'; **grey lines (bottom panel):** six most surprising moments chosen by expert listener.

# Two Pages · Rule based analysis



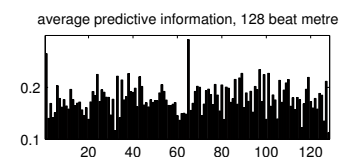
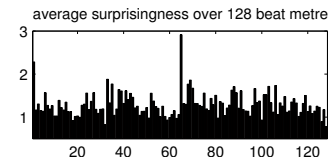
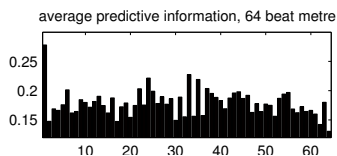
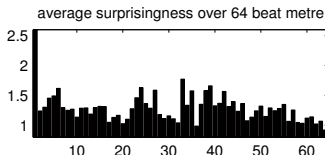
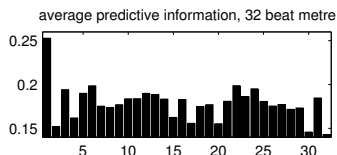
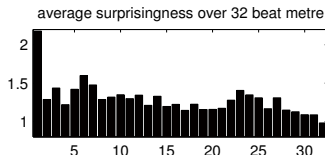
Analysis of *Two Pages* using (top) Cambouropoulos' Local Boundary Detection Model (LBDM) and (bottom) Lerdahl and Jackendoff's grouping preference rule 3a (GPR3a), which is a function of pitch proximity. Both analyses indicate 'boundary strength'.

## Two Pages · Discussion

Correspondence between the information measures and the structure of the piece is quite close. Good agreement between the six ‘most surprising moments’ chosen by expert listener and model information signal.

What appears to be an error in the detection of the major part boundary (between events 5000 and 6000) actually raises a known anomaly in the score, where Glass places the boundary several events before there is any change in the pattern of notes. Alternative analyses of *Two Pages* place the boundary in agreement with peak in our surprisingness signal.

# Gradus · Metrical analysis



# Outline

Expectation and surprise in music

Surprise, entropy and information in random sequences

Markov chains

Application: The Melody Triangle

More process models

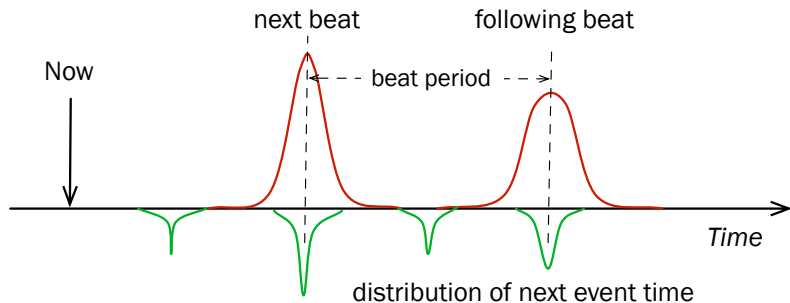
Application: Analysis of minimalist music

**Application: Beat tracking and rhythm**

Summary and conclusions

# Bayesian beat tracker

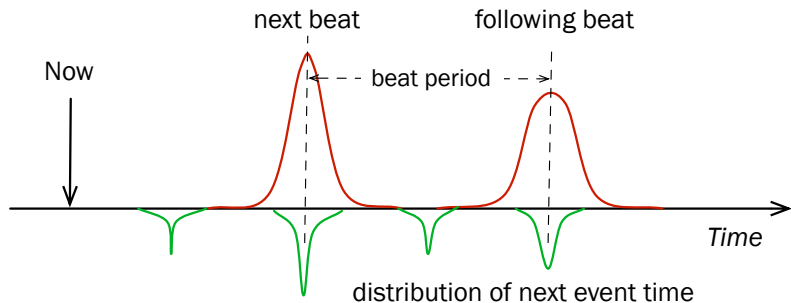
Works by maintaining probabilistic belief state about time of next beat and current tempo.





# Bayesian beat tracker

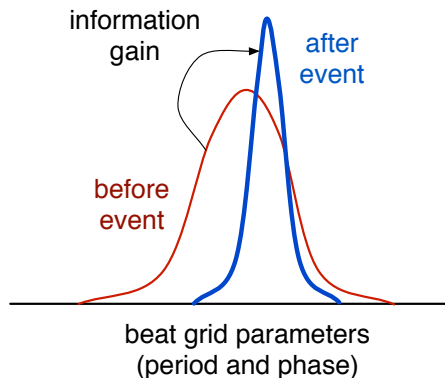
Works by maintaining probabilistic belief state about time of next beat and current tempo.



Receives categorised drum events (kick or snare) from audio analysis front-end.

# Information gain in the beat tracker

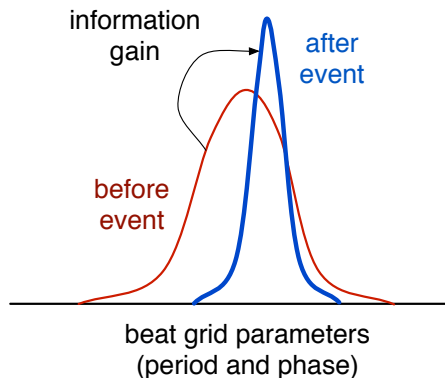
Each event triggers a change in belief state, so we can compute information gain about beat parameters.



# Information gain in the beat tracker

Each event triggers a change in belief state, so we can compute information gain about beat parameters.

Relationship between IIG and IPI means we treat it as a proxy for IPI.



# Analysis of drum patterns

We analysed 17 recordings of drummers, both playing solo or with a band. All patterns in were in 4/4.

- Information tends to arrive at beat times: consequence of structure of model.

# Analysis of drum patterns

We analysed 17 recordings of drummers, both playing solo or with a band. All patterns in were in 4/4.

- Information tends to arrive at beat times: consequence of structure of model.
- Lots of information seems to arrive after drum fills and breaks as the drummer reestablishes the beat.

# Analysis of drum patterns

We analysed 17 recordings of drummers, both playing solo or with a band. All patterns in were in 4/4.

- Information tends to arrive at beat times: consequence of structure of model.
- Lots of information seems to arrive after drum fills and breaks as the drummer reestablishes the beat.
- No consistent pattern of information arrival in relation to metrical structure, so no obvious metrical structure in micro-timing of events. However, still possible that metrical structure might emerge from predictive analysis of drum pattern.

# Outline

Expectation and surprise in music

Surprise, entropy and information in random sequences

Markov chains

Application: The Melody Triangle

More process models

Application: Analysis of minimalist music

Application: Beat tracking and rhythm

**Summary and conclusions**

# Summary

- Dynamic, observer-centric information theory.
- Applicable to any dynamic probabilistic model.
- PIR potentially a measure of complexity.
- Simple analysis for Markov chains and Gaussian processes.
- Applications in music analysis and composition.
- Search for neural correlates is ongoing (that's another talk...).

Thanks!



# Bibliography I

- [Ber71] D. E. Berlyne. *Aesthetics and Psychobiology*. Appleton Century Crofts, New York, 1971.
- [Coh62] J. E. Cohen. Information theory and music. *Behavioral Science*, 7(2):137–163, 1962.
- [Cox46] Richard T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13, 1946.
- [CW95] Darrell Conklin and Ian H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.
- [dF75] Bruno de Finetti. *Theory of Probability*. John Wiley and Sons, New York, 1975.

# Bibliography II

- [ETK02] T. Eerola, P. Toiviainen, and C. L. Krumhansl. Real-time prediction of melodies: Continuous predictability judgments and dynamic models. In C. Stevens, D. Burnham, G. McPherson, E. Schubert, and J. Renwick, editors, *Proceedings of the 7th International Conference on Music Perception and Cognition (ICMPC7)*, Sydney, Australia, 2002. Causal Productions.
- [Han86] Eduard Hanslick. *On the musically beautiful: A contribution towards the revision of the aesthetics of music*. Hackett, Indianapolis, IN, 1854/1986.
- [H095] David Haussler and Manfred Oppen. General bounds on the mutual information between a parameter and  $n$  conditionally independent observations. In *Proceedings of the Seventh Annual ACM Workshop on Computational learning theory (COLT '95)*, pages 402–411, New York, NY, USA, 1995. ACM.

## Bibliography III

- [IB05] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. In *Advances Neural in Information Processing Systems (NIPS 2005)*, volume 19, pages 547–554, Cambridge, MA, 2005. MIT Press.
- [Jay88] Edwin T. Jaynes. How does the brain do plausible reasoning? In G. J. Erickson and C. R. Smith, editors, *Maximum-Entropy and Bayesian Methods in Science and Engineering*. Kluwer Academic, 1988.
- [Mey67] Leonard B. Meyer. *Music, the arts and ideas: Patterns and Predictions in Twentieth-century culture*. University of Chicago Press, 1967.
- [Mol66] Abraham Moles. *Information Theory and Esthetic Perception*. University of Illinois Press, 1966.
- [Nar77] Eugene Narmour. *Beyond Schenkerism*. University of Chicago Press, 1977.

# Bibliography IV

- [Pea05] Marcus T. Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Department of Computing, City University, London, 2005.
- [SJAN99] J. R. Saffran, E. K. Johnson, R. N. Aslin, and E. L. Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52, 1999.
- [VW06] S. Verdú and T Weissman. Erasure entropy. In *IEEE International Symposium on Information Theory (ISIT 2006)*, pages 98–102, 2006.
- [Wun97] W. Wundt. *Outlines of Psychology*. Englemann, Leipzig, 1897.