What/when causal expectation modeling in monophonic pitched melodies and percussive audio

Amaury Hazan, Paul Brossier, Ricard Marxer and Hendrik Purwins

ahazan@iua.upf.edu

Pompeu Fabra University

Barcelona, Spain

http://www.iua.upf.edu/mtg

Outline

- Goals
- Background
- System Design
- Evaluation
- Future Work
- Conclusions

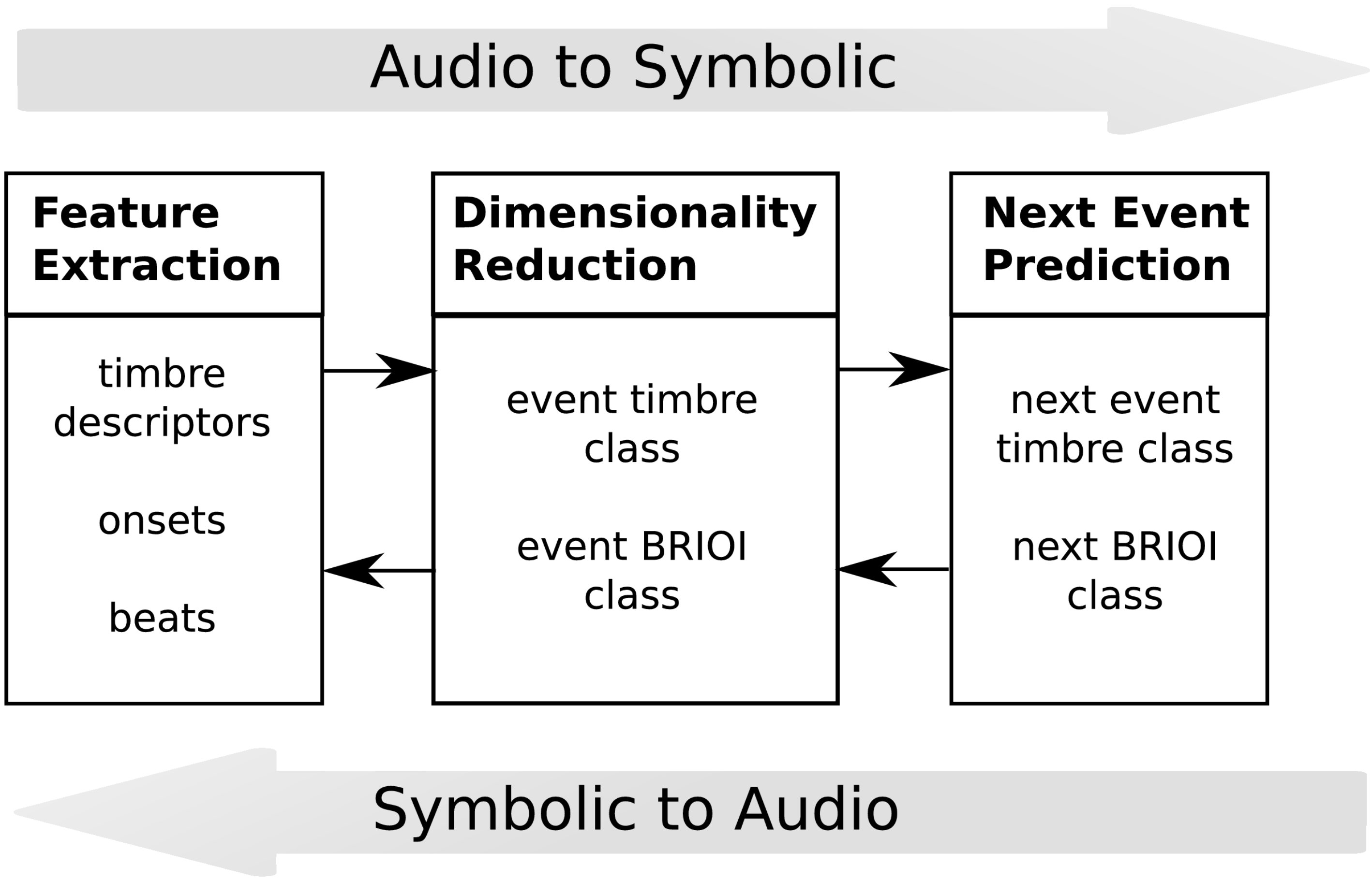
Goals

- Build a system for producing musical expectations based on the observation of musical audio
- The system has to be unsupervised and causal to respect cognitive constraints
  - this enables to study the effect of exposure
    A use case for musical sequence learning models
- what/when expectation task
- As a first step, we focus on constant tempo musical patterns
  - Drum loops
  - Monophonic Pitched Melodies

Background

- Prediction-driven listening [Ellis, Abdallah]
- Symbolic pitch sequences learning systems
  - [Todd, Mozer, Tillmann, Eck & Schmidhuber, Pearce & Wiggins]
- Bayesian score following and accompaniment of audio signals
  - [Raphael, Cemgil, Orio]
- Unsupervised learning and concatenative synthesis
  - [Schwarz, Jehan]
- Improvisation systems
  - [Pachet, Assayag, Dubnov, Cont]

System Design

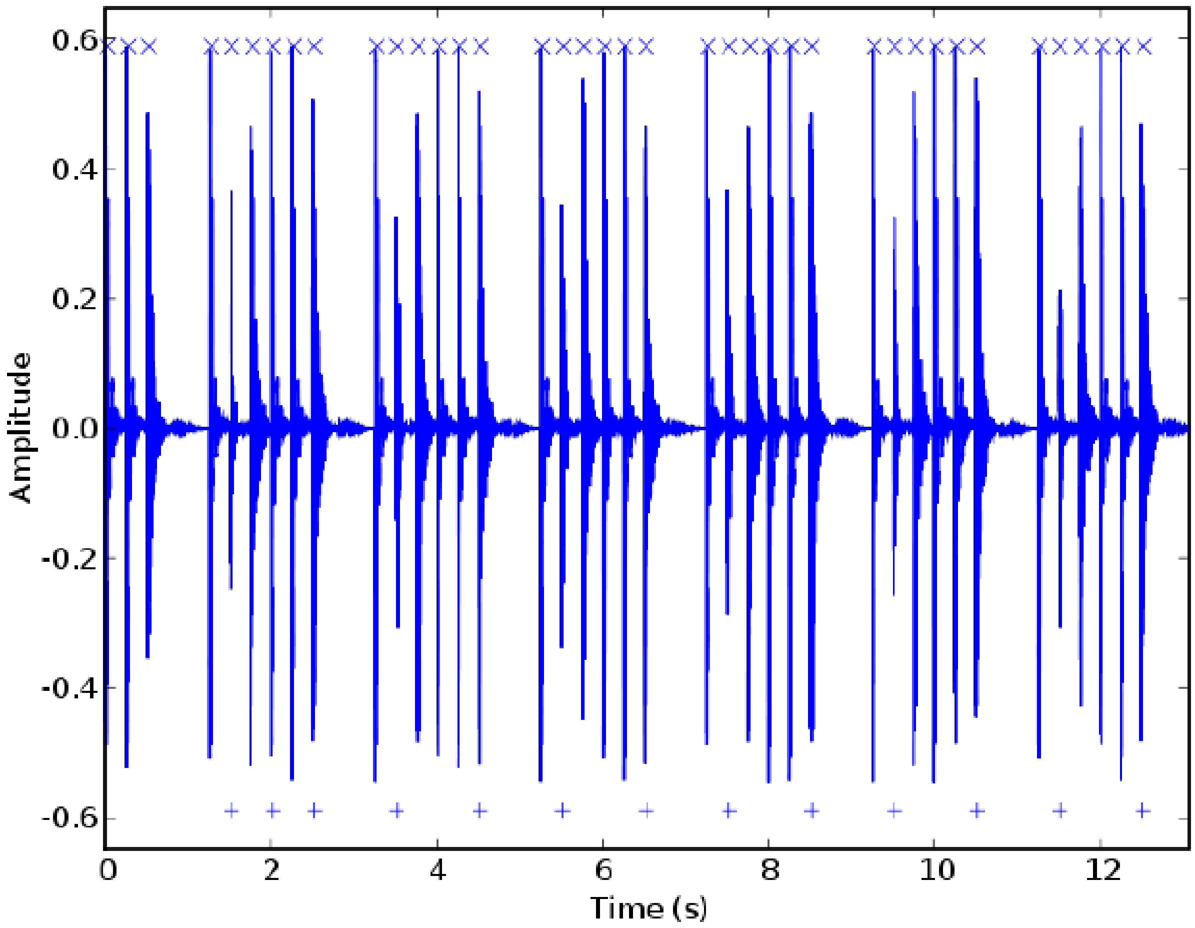


Feature Extraction Module

Extracts on-the-fly the following features:

- Beats [Davis et al. 2005]
- Onsets (High frequency content based)
  - meter description is not explicitly modeled
- Timbre Descriptors
  - rough spectral shape (ZCR, SC)
  - MFCC
  - pitch (YIN-FFT, [Brossier 2004])
- Timbre descriptors can be computed on the onset frame (fast), or averaged over the IOI region

Feature Extraction Module: output



Onsets

Beats

Dimensionality Reduction Module

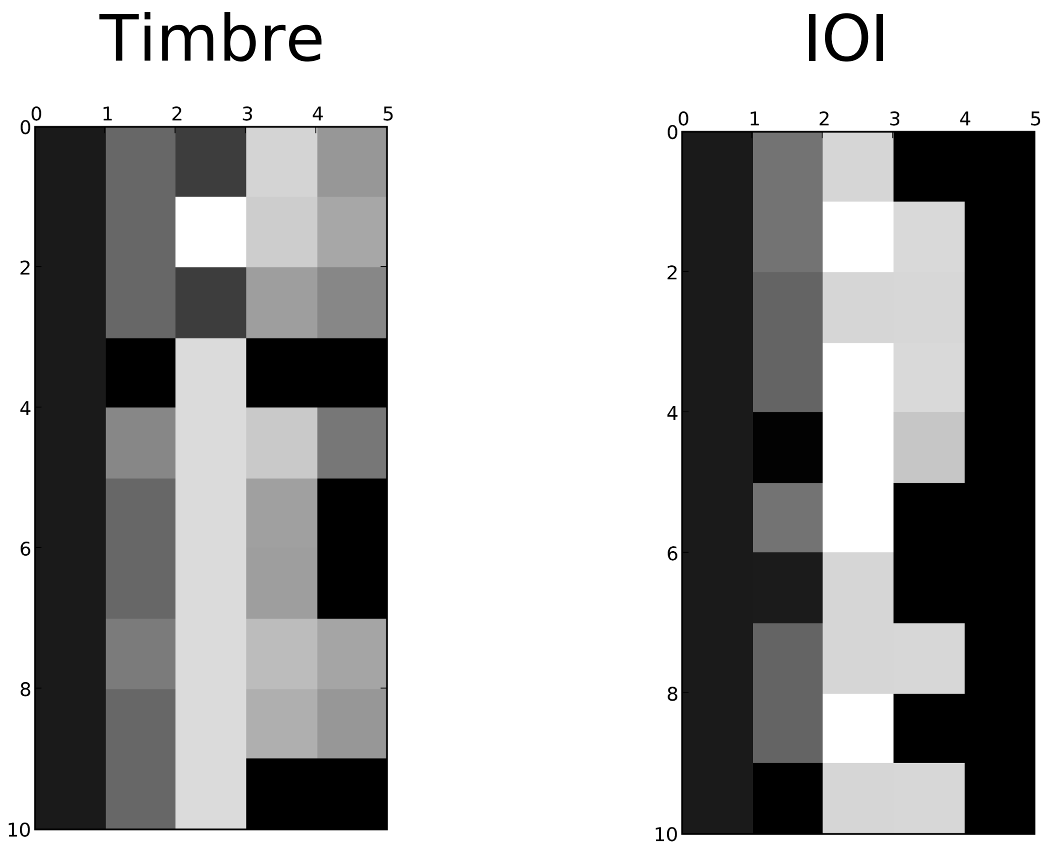
- On-line unsupervised clustering to create symbols for both temporal and timbre features

- Prior to this, we perform a *bootstrap* step
  - Accumulates timbre features and beat-relative IOI
  - We normalize the timbre
  - GMM+EM grid, choose best number of components

Bootstrap GMM+EM grid

- Create GMM grid
  - each row is an independent run
  - for each row, each column is associated with a GMM with number of components= 1, 2, ... , K_{max}
- Fit each GMM using EM
  - Simple regularization procedure to avoid excessively low variances.
- For each model, compute information criterion
  - BIC, AIC, AICc
- Get median of best K over each row

GMM Grid: Example (drums)



Running state: Online K-Means

- For each incoming point
  - Cluster assignment

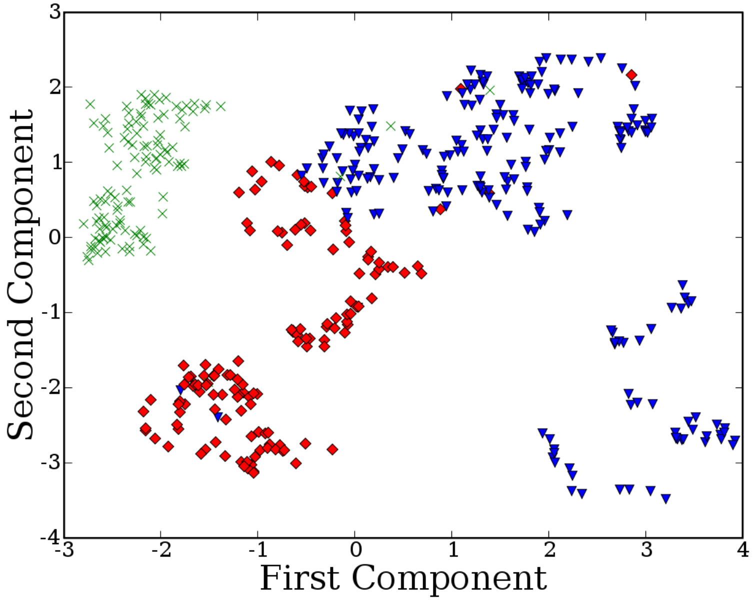
$$C(x) = \mathrm{argmin}_{1<j<K} ||x - \mu_j||^2$$

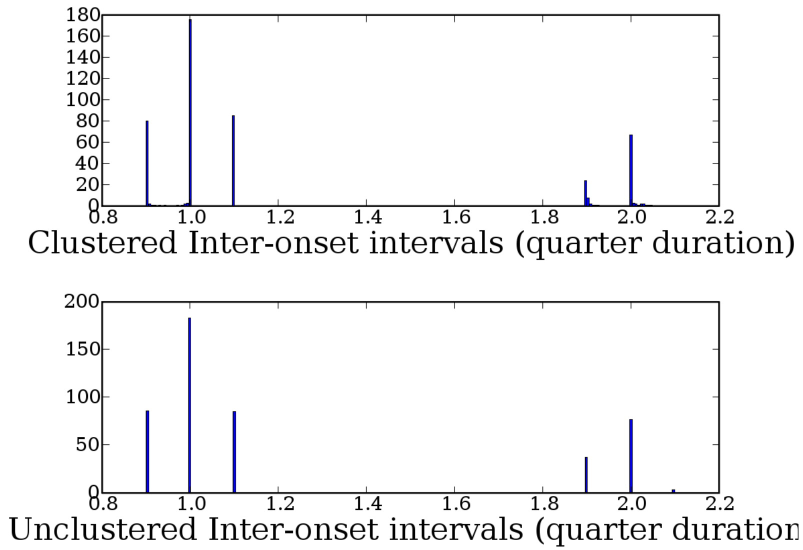
  - Cluster mean update

$$\Delta\mu_j = \eta(x - \mu_j)$$

  - η is the learning rate

Dimensionality Reduction Module: Output (drums)

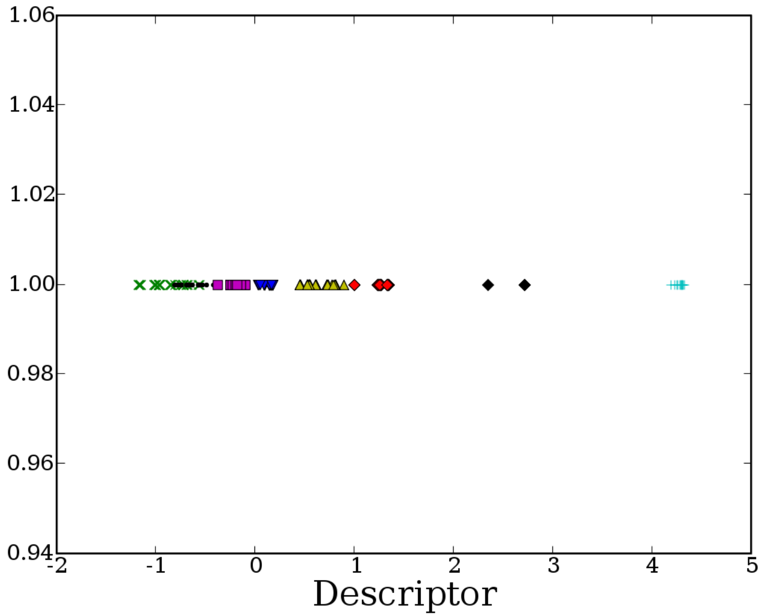


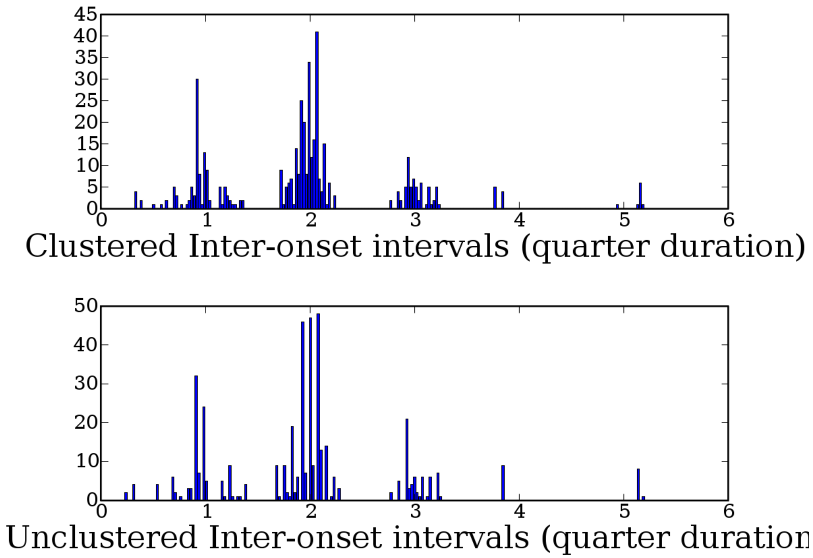


Timbre Symbols

Time Symbols

Dimensionality Reduction Module: Output (sung melody)

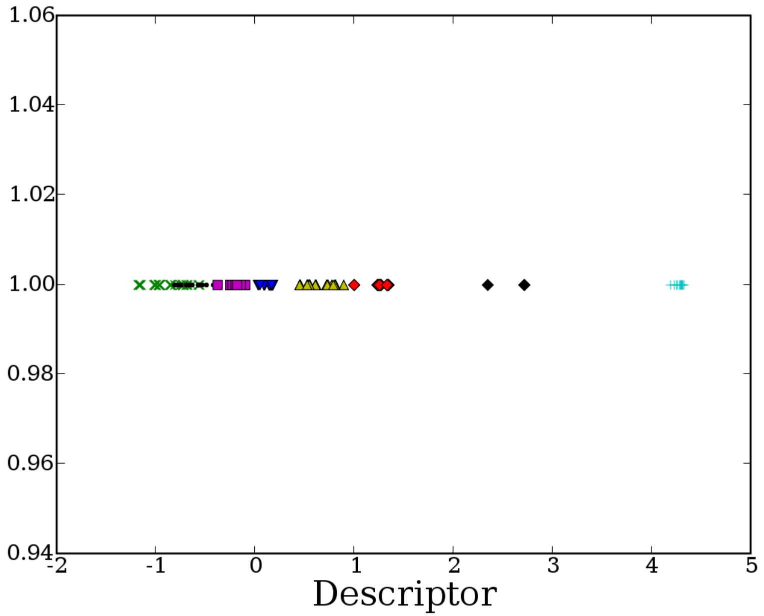


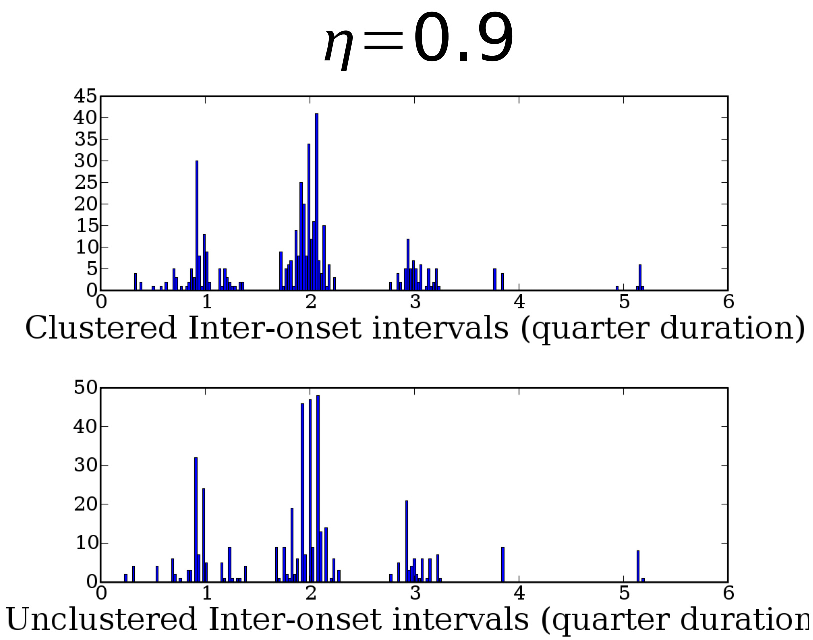


Timbre Symbols

Time Symbols

Dimensionality Reduction Module: Output (sung melody)

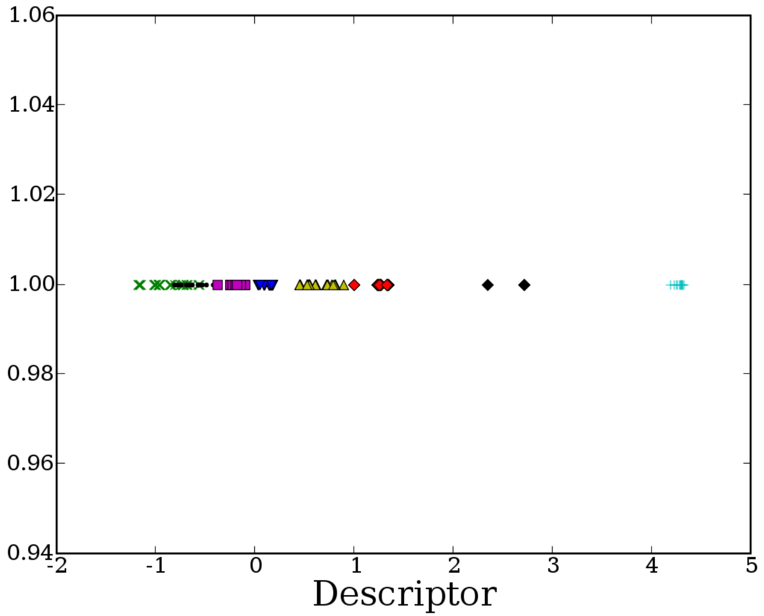


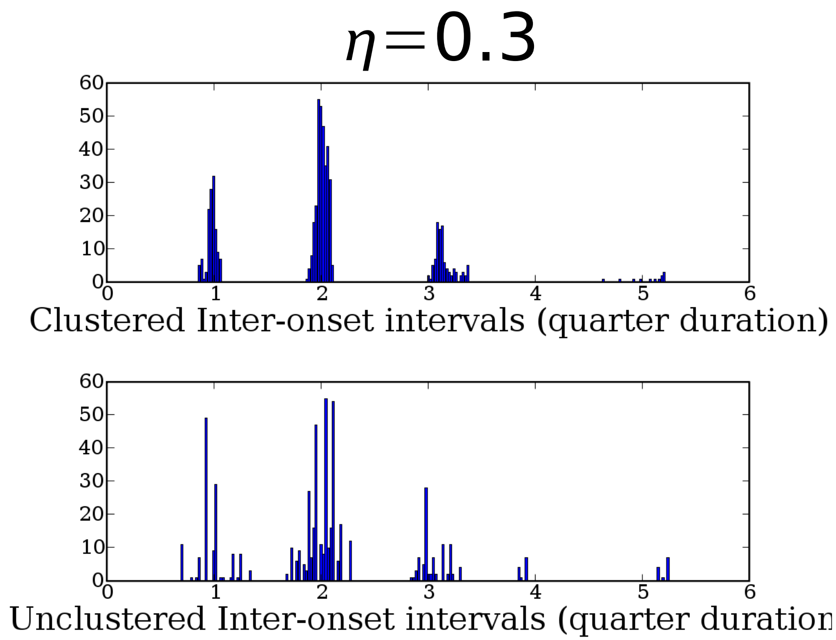


Timbre Symbols

Time Symbols

Dimensionality Reduction Module: Output (sung melody)





Timbre Symbols

Time Symbols

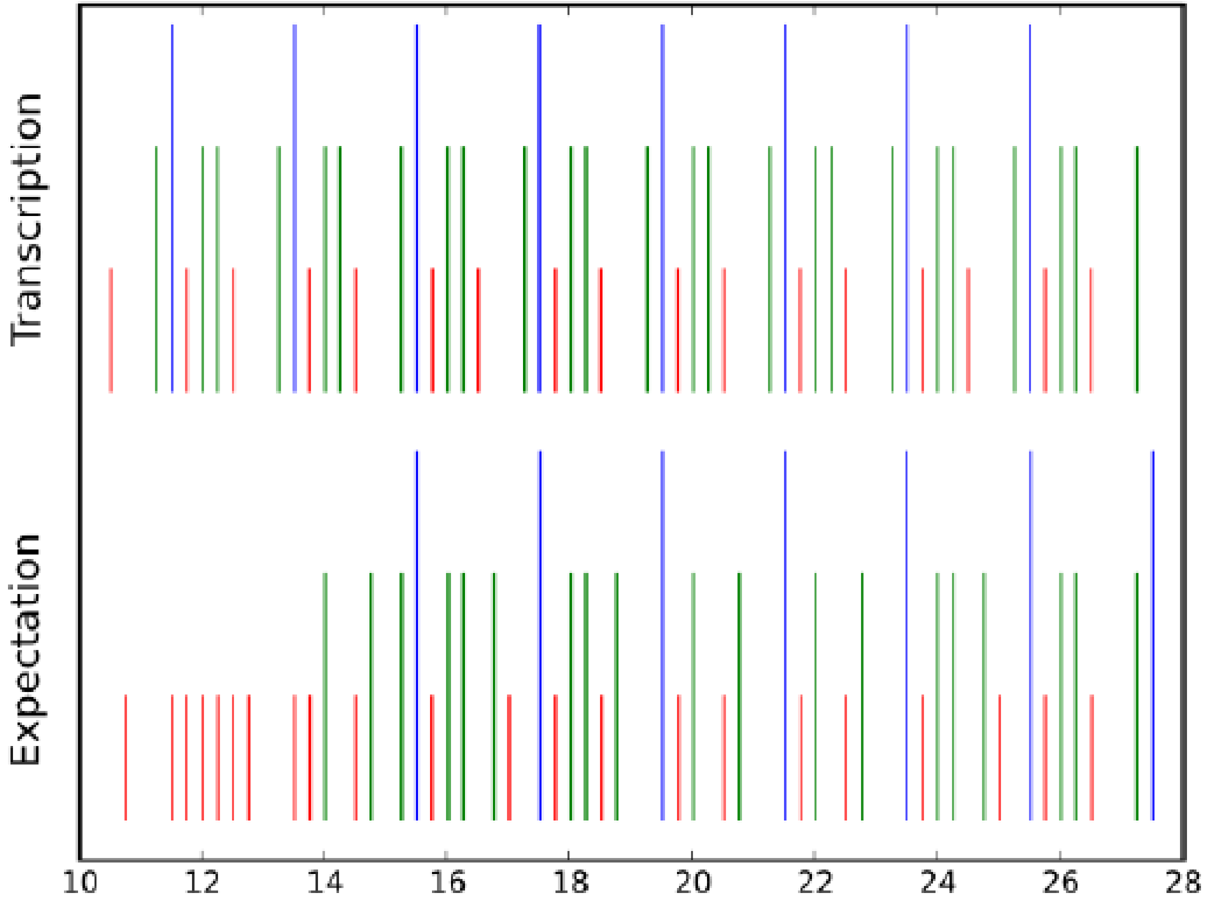
Prediction By Partial Match [Cleary&Witten, Pearce&Wiggins]

- probability of next symbol given context sequence

$$p(e_i|e_{(i-n)+1}^{i-1}) = \begin{cases} \alpha(e_i|e_{(i-n)+1}^{i-1}) & \text{if } c(e_i|e_{(i-n)+1}^{i-1}) > 0 \\ \gamma(e_{(i-n)+1}^{i-1})p(e_i|e_{(i-n)+2}^{i-1}) & \text{if } c(e_i|e_{(i-n)+1}^{i-1}) = 0 \end{cases}$$

- α is computed based on the counts of a given symbol after the observed context
- γ controls the recursive *backoff*
  - enables to integrate predictions based on lower order contexts when needed
  - if symbol never seen before in any context: uniform distribution

Next Event Prediction Module: output



Evaluation

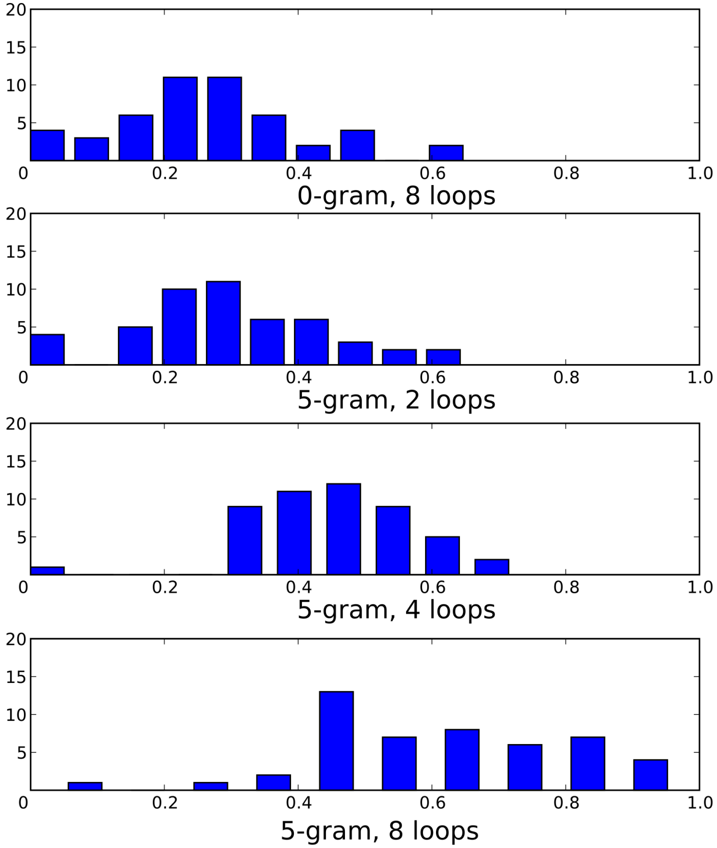
- Listening to looped patterns
- We compute a *weighted F-Measure* to compare transcription and expectation
  - weights needed because unused clusters can appear

$$WFM = \sum_{i=1}^{K_t} w_i F_i$$

- The systems performs twice better than using random predictors

Evaluation: Drums and Sung Melody

- Drums: WFM histogram



- Drums: Depending of descriptor set

ZCR, SC	MFCC
0.60 (3.22, 2.33)	*0.69 (1.50, 2.42)*

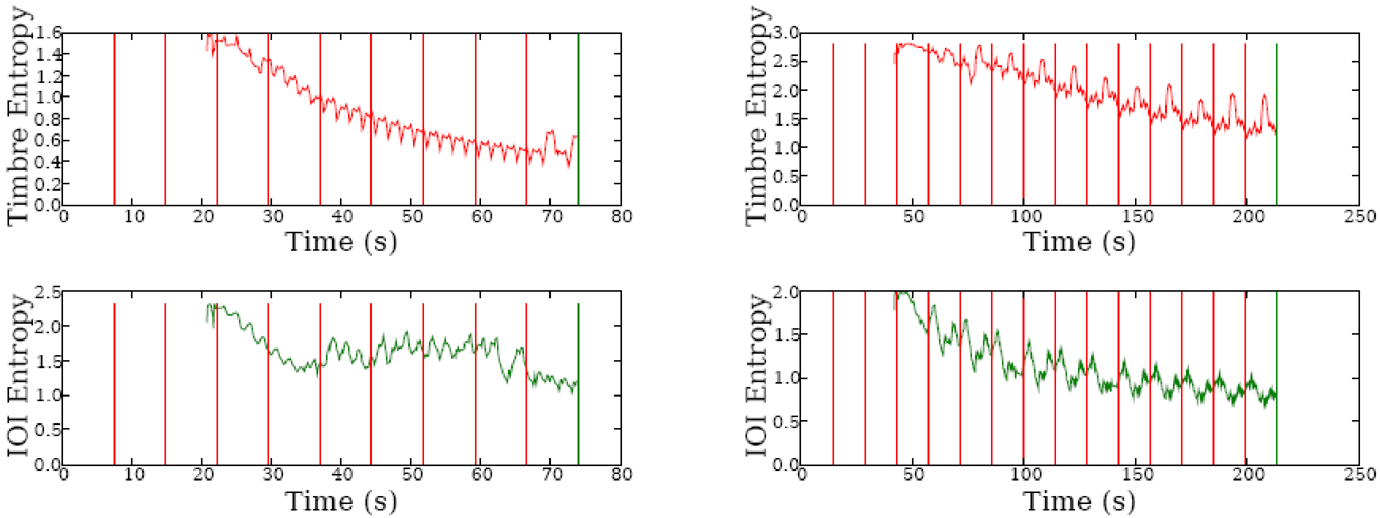
- Sung Melodies

Excerpt	Folk.1	Folk.2	Folk.3
Exp.2	0.08 (7, 4)	0.32 (5, 3)	0.24 (5, 3)
Exp.4	0.22 (6, 4)	0.34 (6, 3)	0.25 (5, 5)
Exp.8	0.64 (7, 5)	0.53 (7, 3)	0.37 (5, 2)

Expectation Entropy

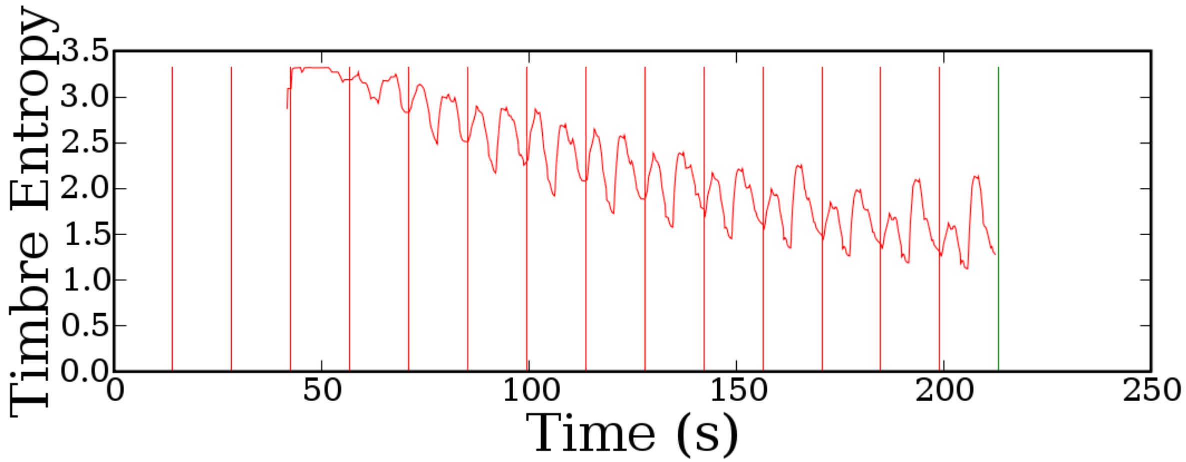
- PPM gives posterior distribution over possible next symbol
- We compute expectation entropy

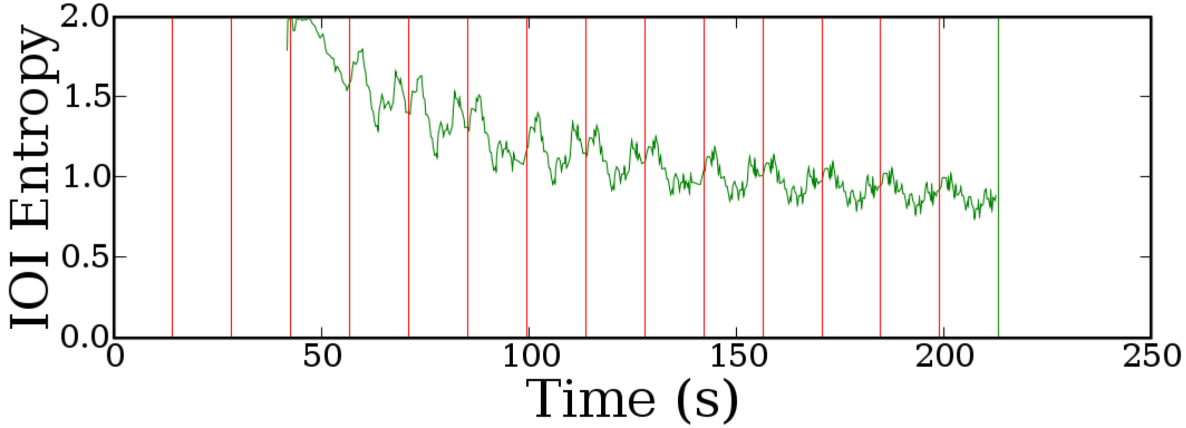
$$H(p) = -\sum_{K} p(e_i) log_2 p(e_i)$$



(a) Commercial jungle drum pattern (b) Monophonic sung melody

Expectation Entropy





Concatenative Synthesis

- [Schwarz, Jehan]
- For each predicted timbre symbol, concatenate in output stream a prototypical audio slice of this symbol having the predicted IOI length
- Examples
  - Drums
  - Melody

Discussion

- Bootstrapping is cheating
  - Cheat more: define a timbre and timing distance based on a whole collection
  - Bootstrap step run in parallel instead of once [Marxer]
    - allows to track variations in the attended signal
    - clusters can appear and disappear
- A less discrete system
  - from hard to soft cluster assignment [McKay]
  - work with transient regions instead of crisp onsets

Summary

- An on-line and unsupervised system for computing expectation in audio signals
- Can be applied to different kinds of monophonic musical signals
- But the sequential prediction system is purely symbolic and markov-chain based
- Expectation entropy may be used to mark temporal cues in the attended signal

Thanks

PPM

$$\gamma(e_i|e_{(i-n)+1}^{i-1}) = \frac{t(e_{(i-n)+1}^{i-1})}{\sum_K c(e_{(i-n)+1}^{i-1}) + t(e_{(i-n)+1}^{i-1})}$$

$$\alpha(e_i|e_{(i-n)+1}^{i-1}) = \frac{c(e_i|e_{(i-n)+1}^{i-1})}{\sum_K c(e|e_{(i-n)+1}^{i-1}) + t(e_{(i-n)+1}^{i-1})}$$