Multiple-F0 tracking based on a high-order HMM model

Wei-Chen Chang
Alvin W.Y. Su
Chunghsin Yeh
Axel Roebel
Xavier Rodet

Digital Audio Effects 2008
DAFx08

SCREAM (Studio of Computer REseArch on Music and Multimedia)

- National Cheng Kung University, Taiwan
- **Wei-Chen Chang** (Mediatek Inc.)

  audio codec, physical modeling of acoustic instruments, machine learning, anaysis and synthesis of musical signals

  visiting researcher at IRCAM (2007-2008)
- **Prof. Alvin W.Y. Su** (director of Campus Information System Group, director of SCREAM)

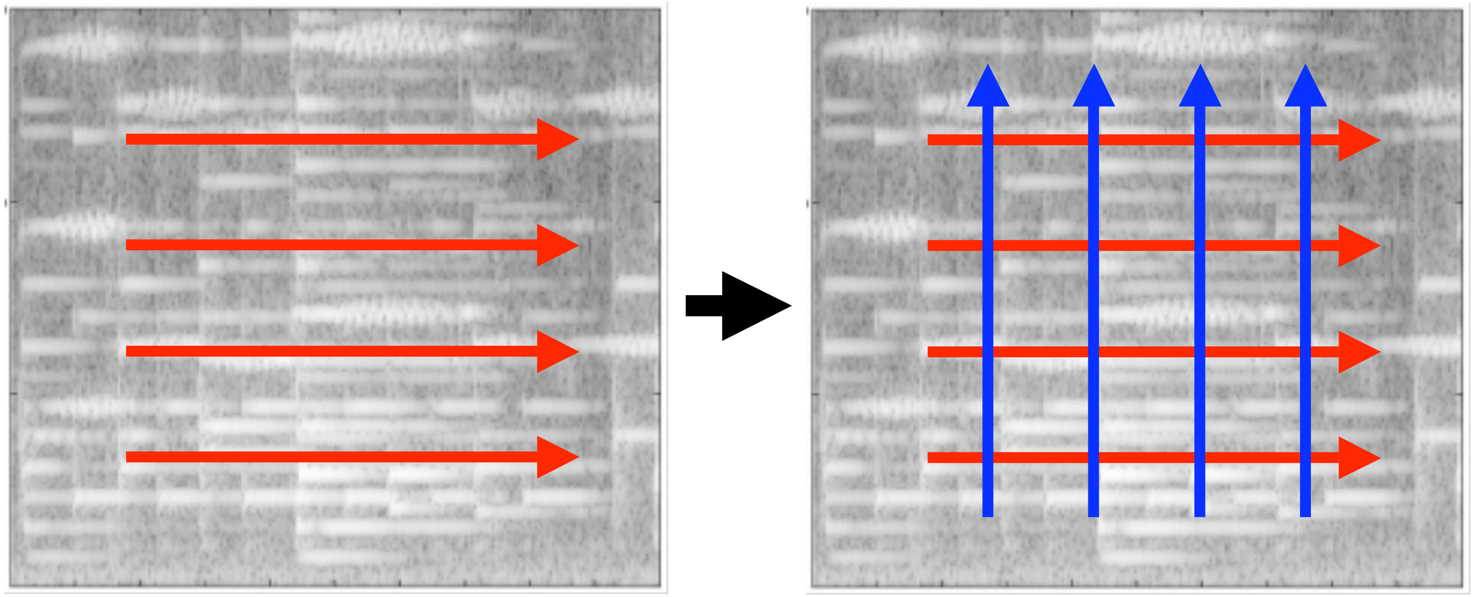
  digital audio/video signal processing, physical modeling of acoustic instruments (CCRMA), multimedia data compression, P2P multimedia streaming systems, embedded systems, VLSI signal processor design and ESL (Electronic System Level) tool design
- WOCMAT (Workshop on Computer Music and Audio Technology)

Outline

- Introduction
  - System overview
- Tracking phase
  - Forward propagation
  - Iterative backward tracking
- Pruning phase
  - Estimate the number of source streams
- Evaluation
- Conclusion

Two main approaches

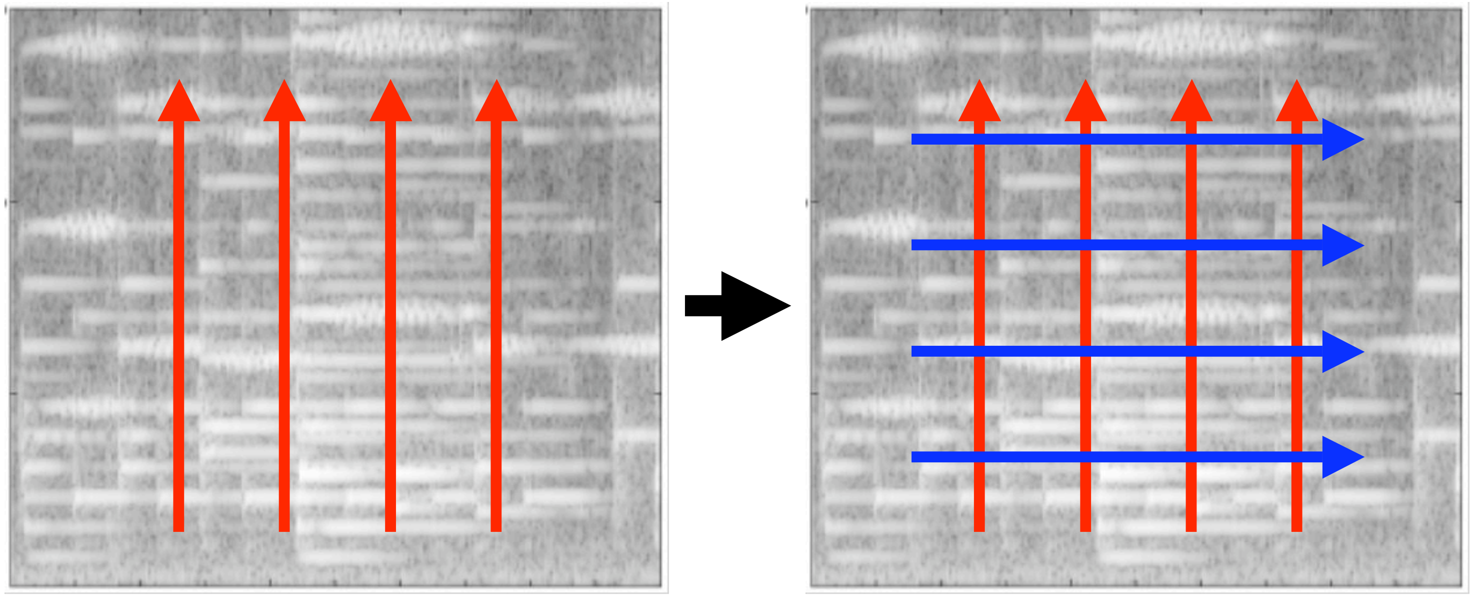
1. Tracking followed by clustering (TfC)



[Mellinger 91],[Martin 96], [Sterian 99], [Lagrange 07]

Two main approaches

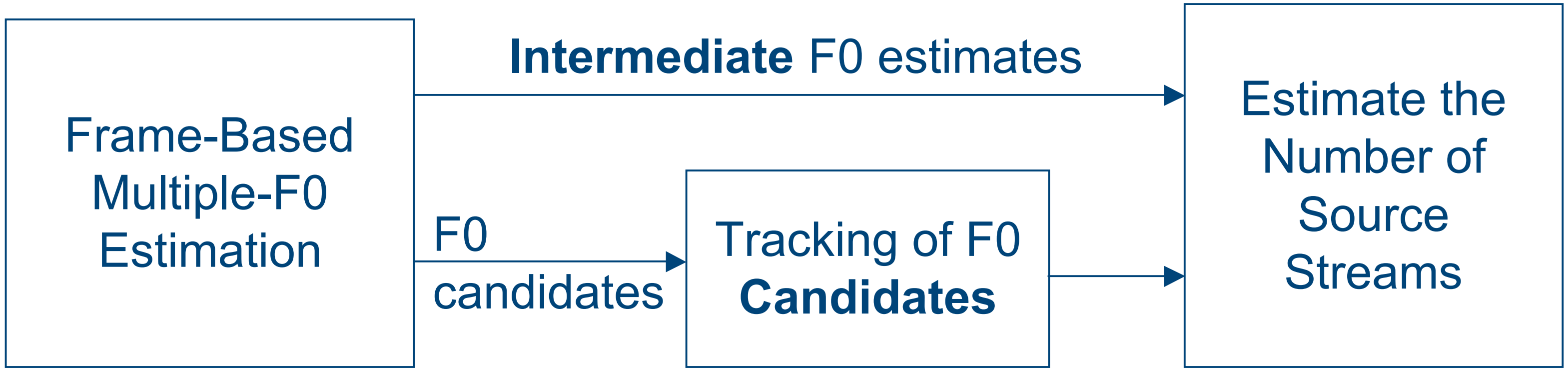
2. Clustering followed by tracking (CfT)



[Wu 03],[Ryynanen 05]

Overview of the proposed system

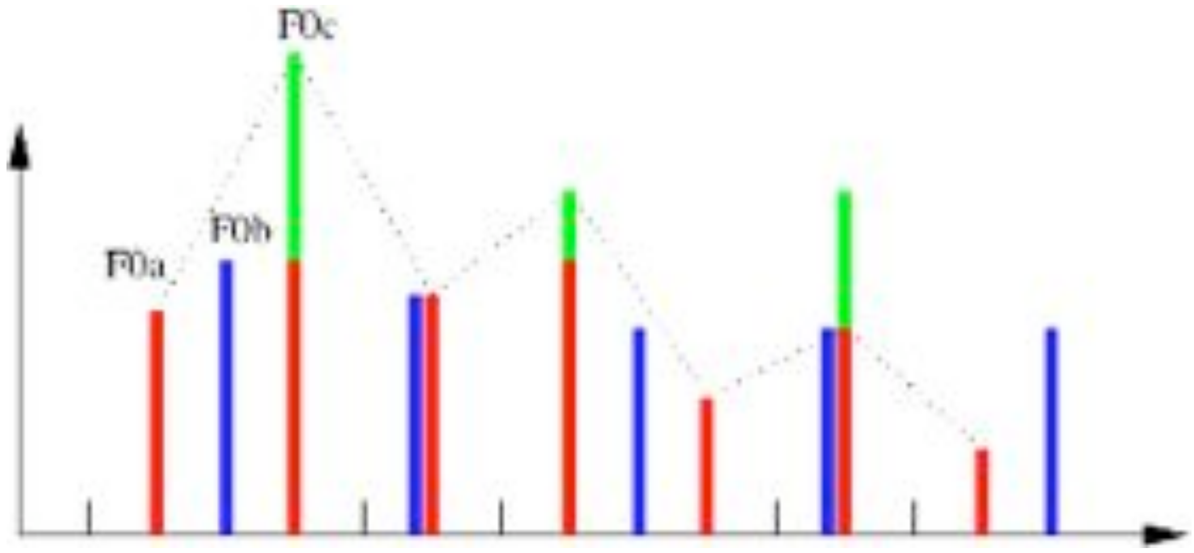
- We follow the CfT approach


Frame-Based Multiple-F0 Estimation
Intermediate F0 estimates
F0 candidates
Tracking of F0 **Candidates**
Estimate the Number of Source Streams

- Frame-based F0 estimation + polyphony inference
- Connection of F0 estimates is fragmentary
- Candidate trajectories are more complete

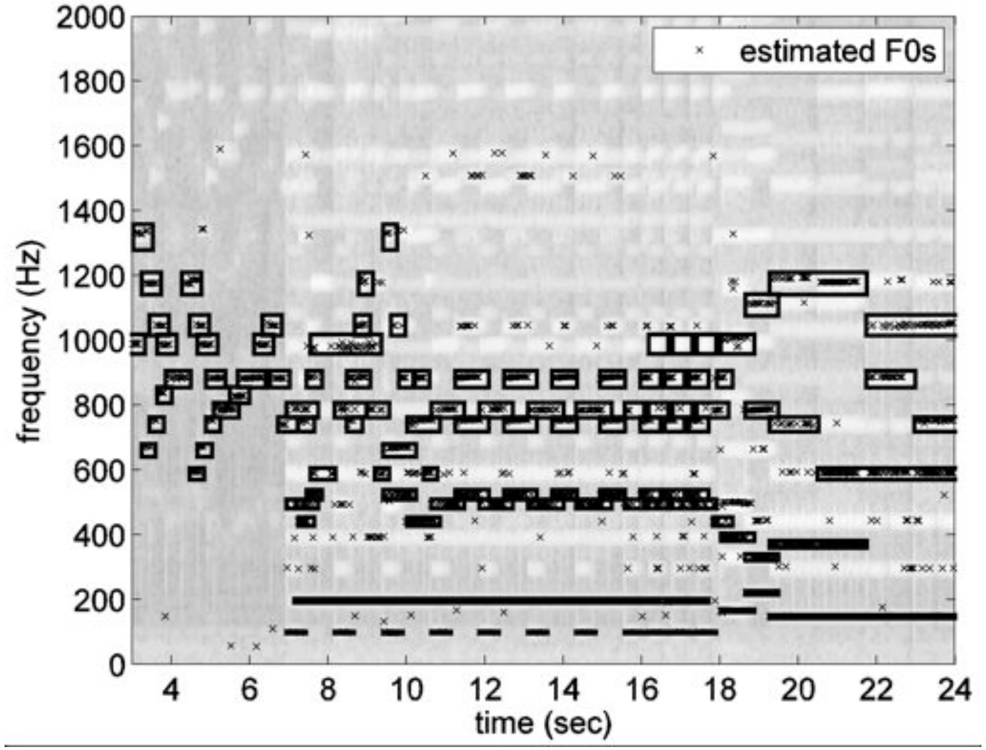
Frame-based polyphony inference

- Two groups of F0s
  - NHRF0 (non-harmonically related F0s): noise level
  - HRF0 (harmonically related F0s): spectral smoothness



Frame-based polyphony inference (cont.)

- Good accuracy



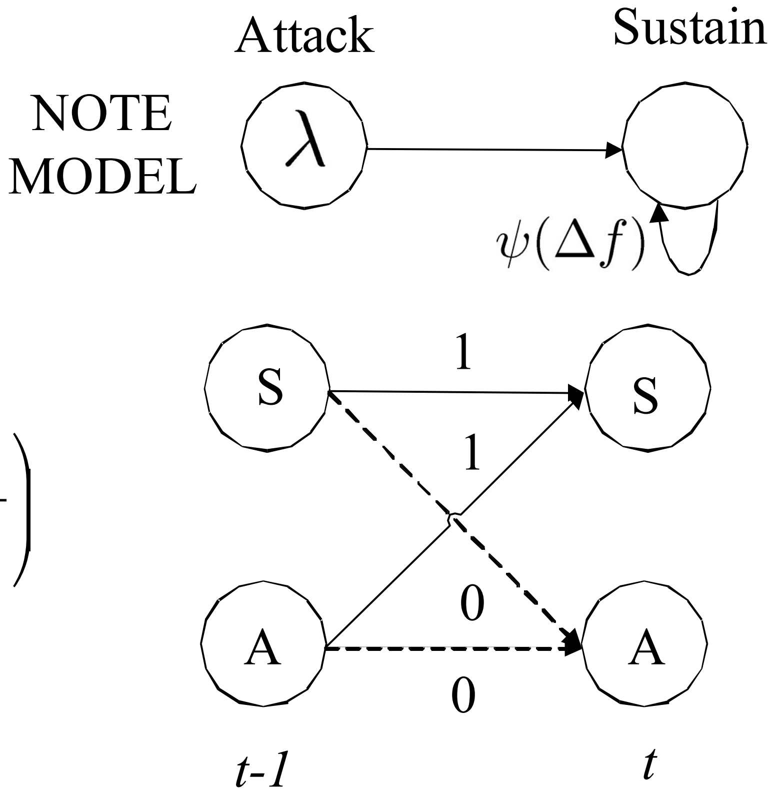
Tracking of F0 candidates using a high-order HMM

- Attack state
  - Attack probability

    λ : parameter
- Sustain state
  - Sustain probability

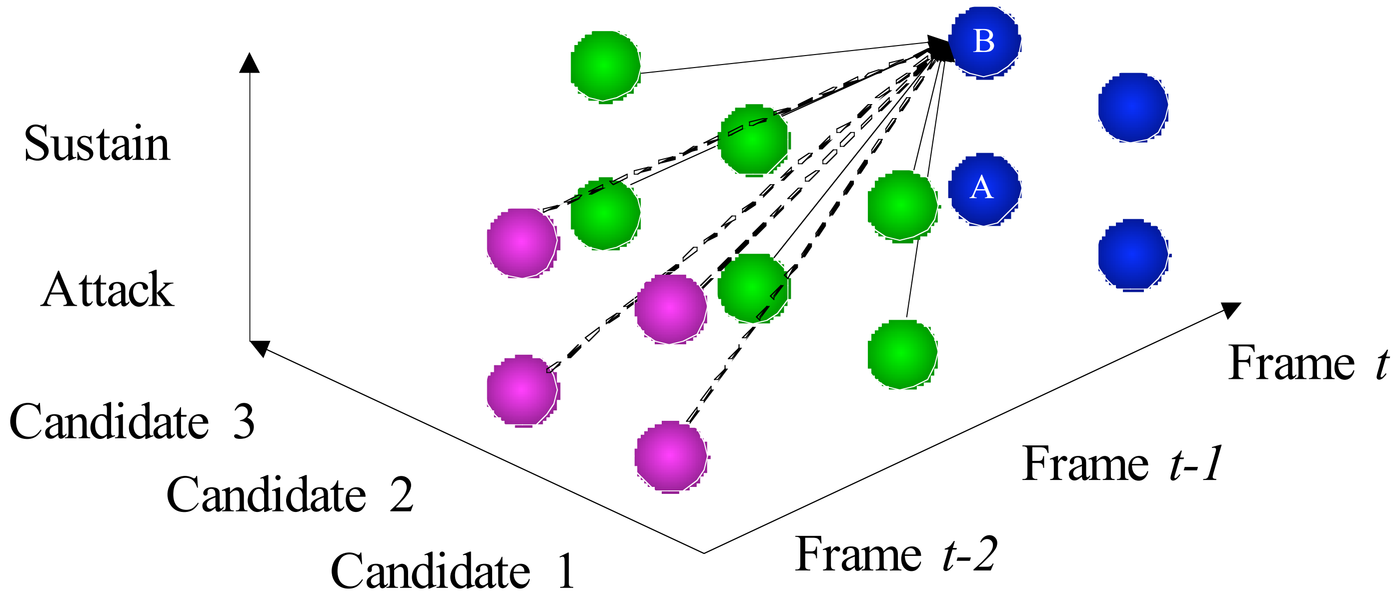
$$\psi(\Delta f) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\Delta f^2}{2\sigma^2}\right)$$

- Transition matrix



Tracking of F0 candidates using a high-order HMM (Cont.)

- HMM’s order is a parameter.



- Forward-backward dynamic programming scheme

Forward propagation of connection weights

- From n(t-d,k,p) to n(t,c,q) :

Weighting parameter

$$\gamma\left(t,c,q \mid t-d,k,p\right) = \alpha\cdot\omega(d)\cdot\Gamma\left(t-d,k,p\right) + \left(1-\alpha\right)\cdot\psi\left(\Delta f\right)$$

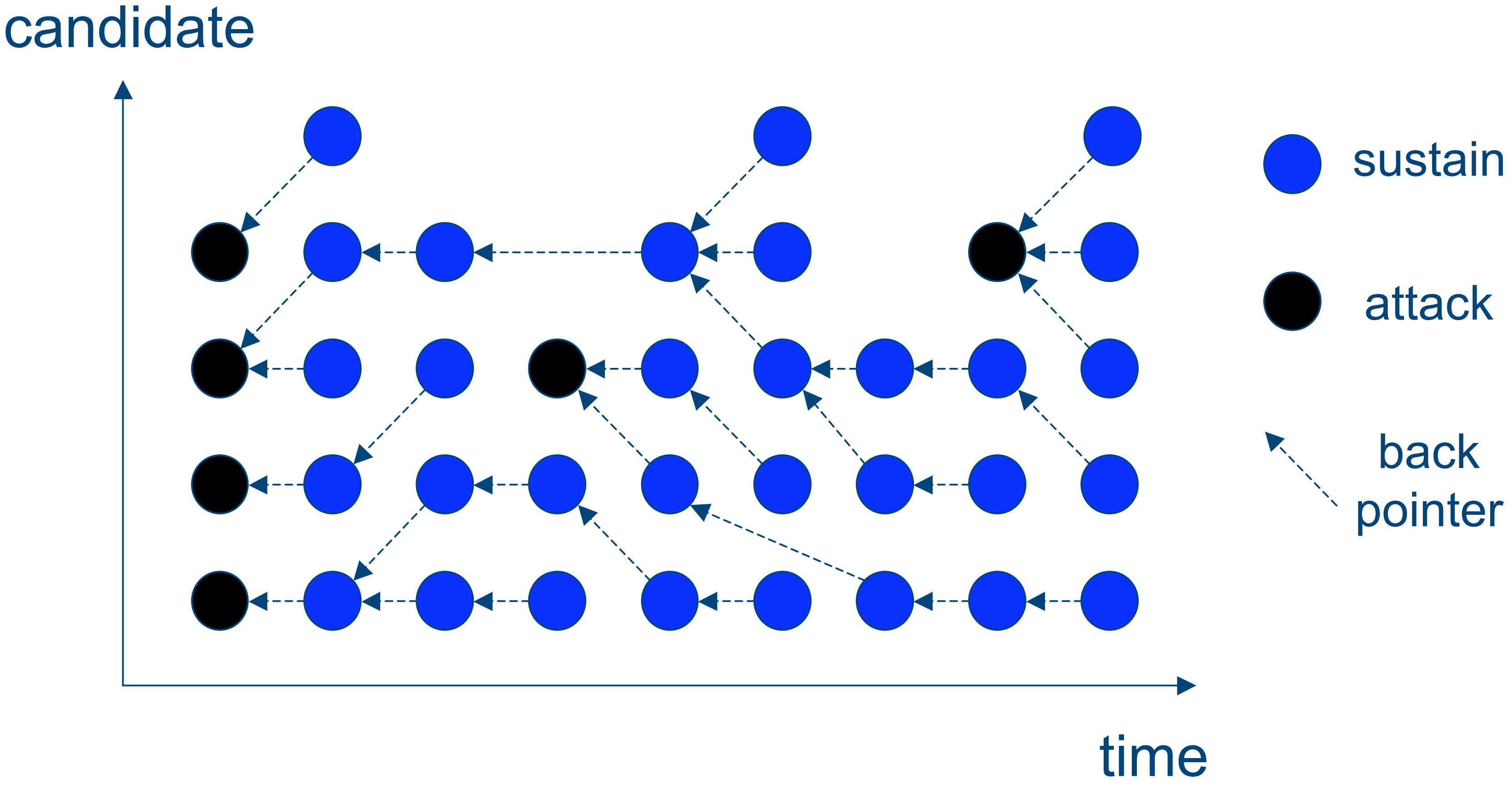
- Back pointer

Decay weighting: $\omega(d) = \frac{1}{d^{s}}$

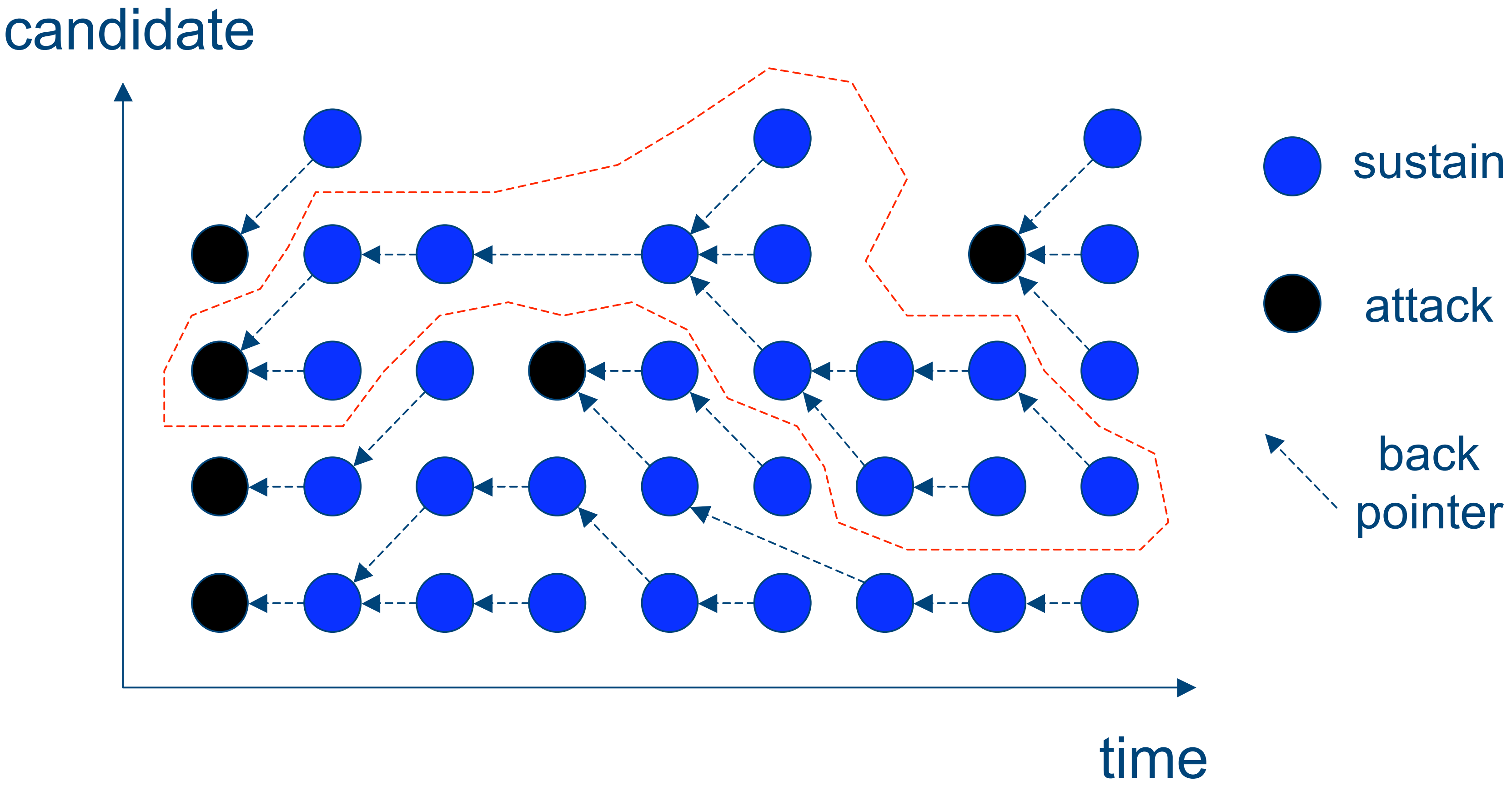
$$I_{\max}\left(t,c,q\right) = \arg\max_{d,k,p} \gamma\left(t,c,q \mid t-d,k,p\right)$$

update $\Gamma\left(t,c,q\right) = \gamma\left(t,c,q \mid I_{\max}\left(t,c,q\right)\right)$

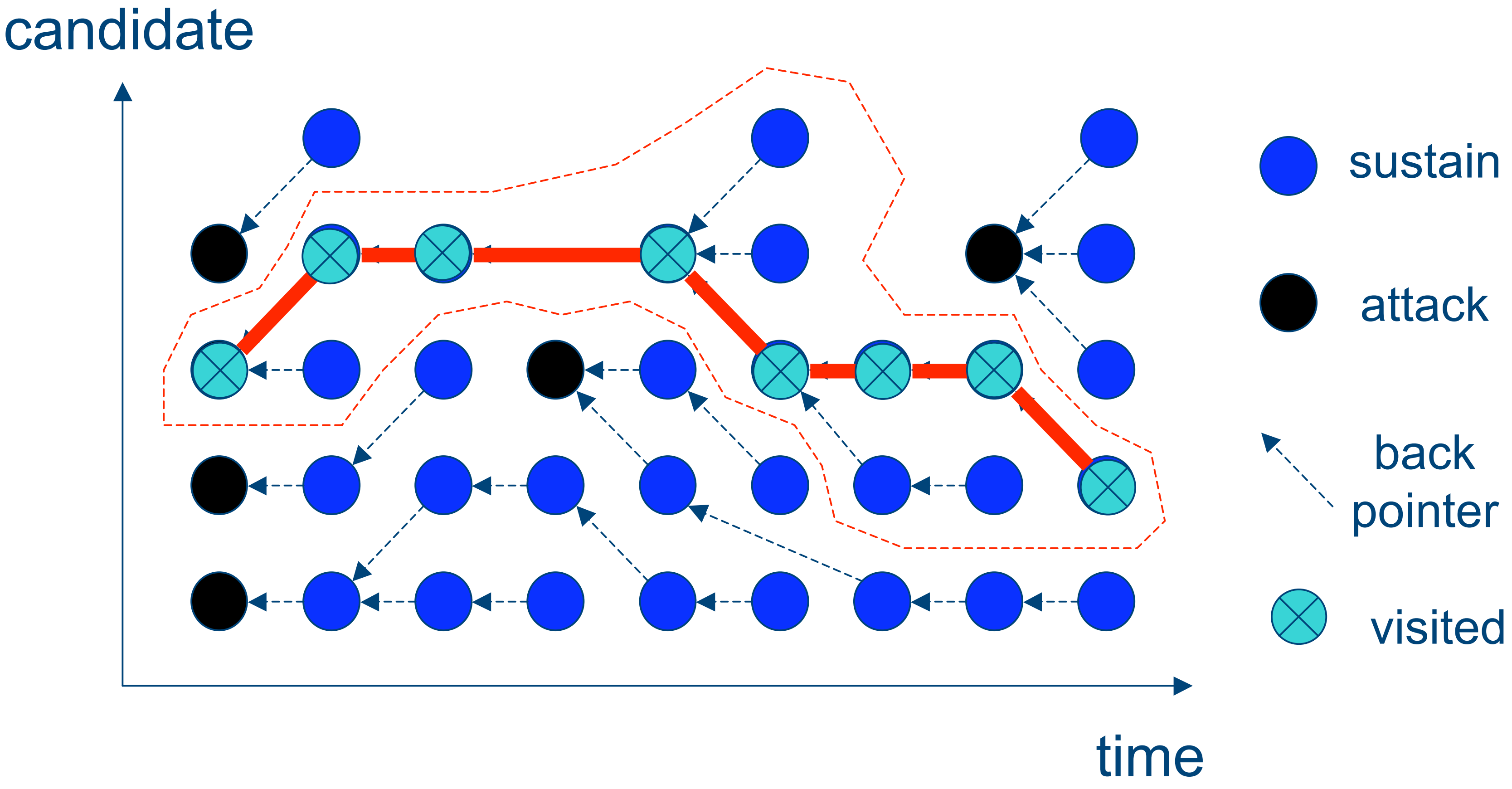
Iterative backward tracking



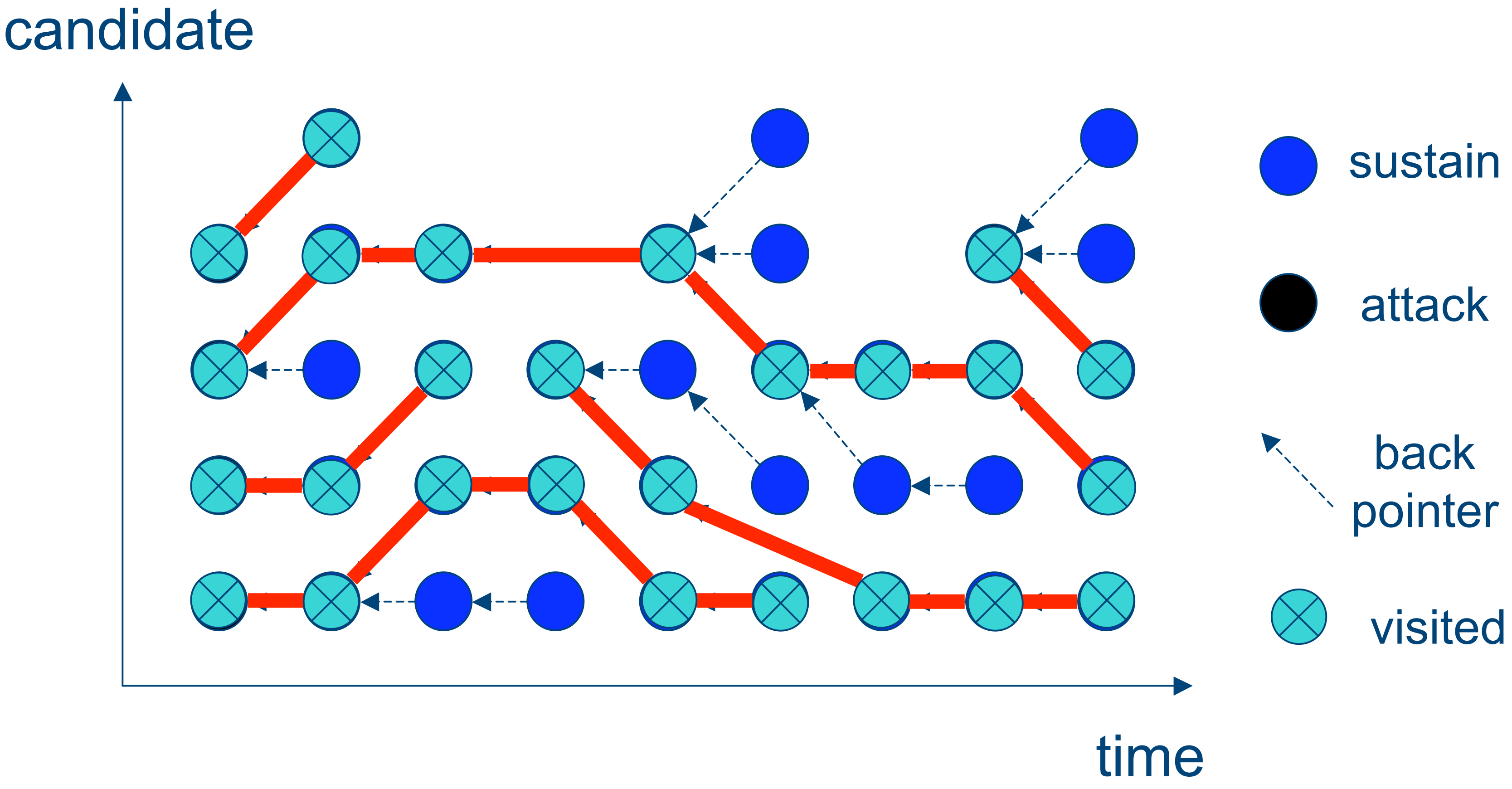
Iterative backward tracking



Iterative backward tracking



Iterative backward tracking



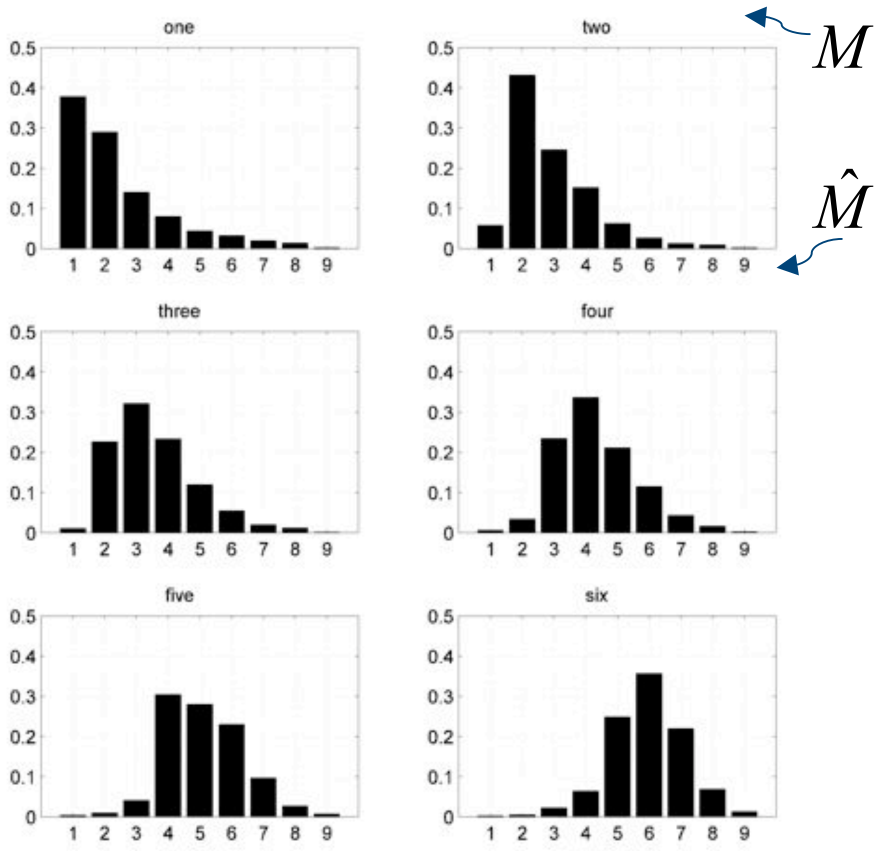
Estimate the number of source streams

- The inferred polyphony estimated by frame-based F0 estimator provides the *reference polyphony* M
- The pruning of the candidate trajectories provides the *estimated polyphony* $\hat{M}$
- The problem is to maximize the log likelihood of $p(M|\hat{M})$ for all observed frames.

Estimate the number of source streams (Cont.)

- By investigating **inference likelihood** of frame-based estimator, it can be modeled by the probability of the polyphony error

$$\Delta M = |M - \hat{M}|$$

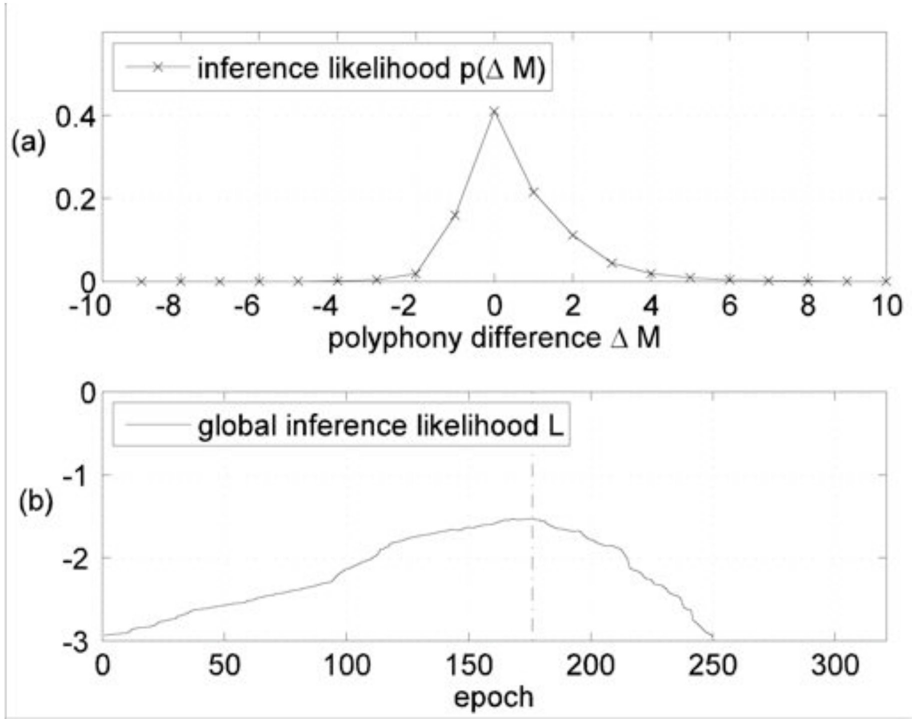


Estimate the number of source streams - pruning

- Log likelihood of the current set of candidate trajectories

$$L = \sum_{t=1}^{T} \log p_t(\Delta M)$$

- Iteratively pruning: the solution is sensitive to the pruning order



Estimate the number of source streams - pruning order

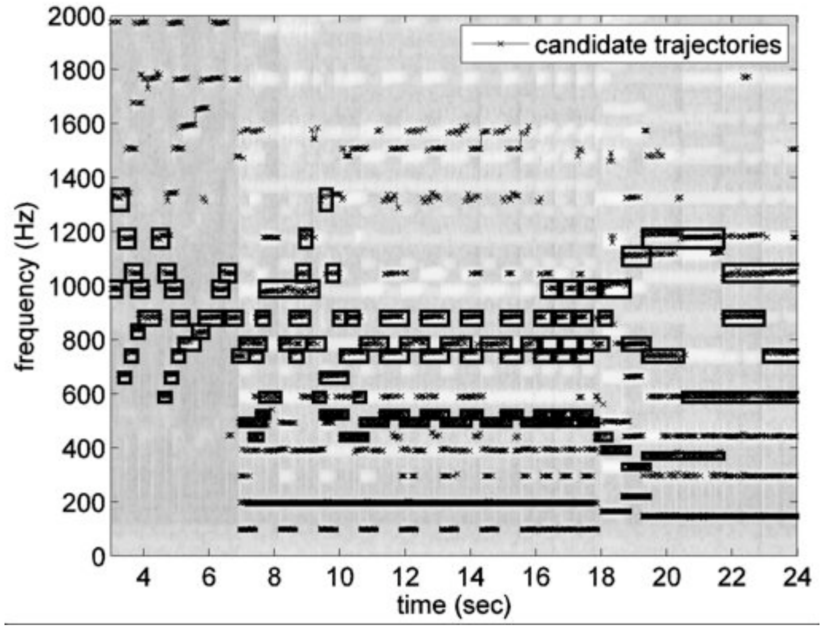
- Accordance ratio
  - related to one single trajectory, T_k .

$$R = \frac{\text{number of intermediate F0 estimates in } T_k}{\text{length of } T_k}$$

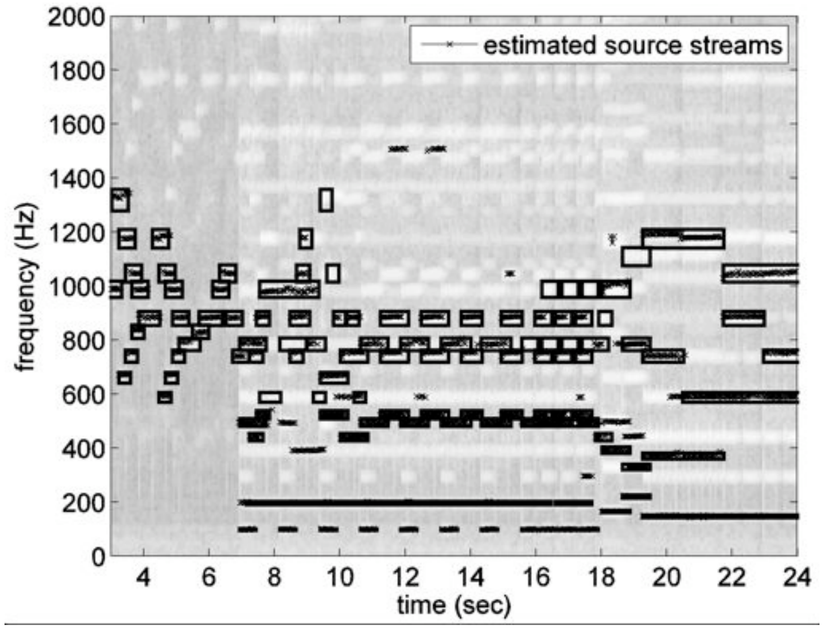
  - determines the pruning order.

Estimate the number of source streams - example

Candidate trajectories



Final estimates

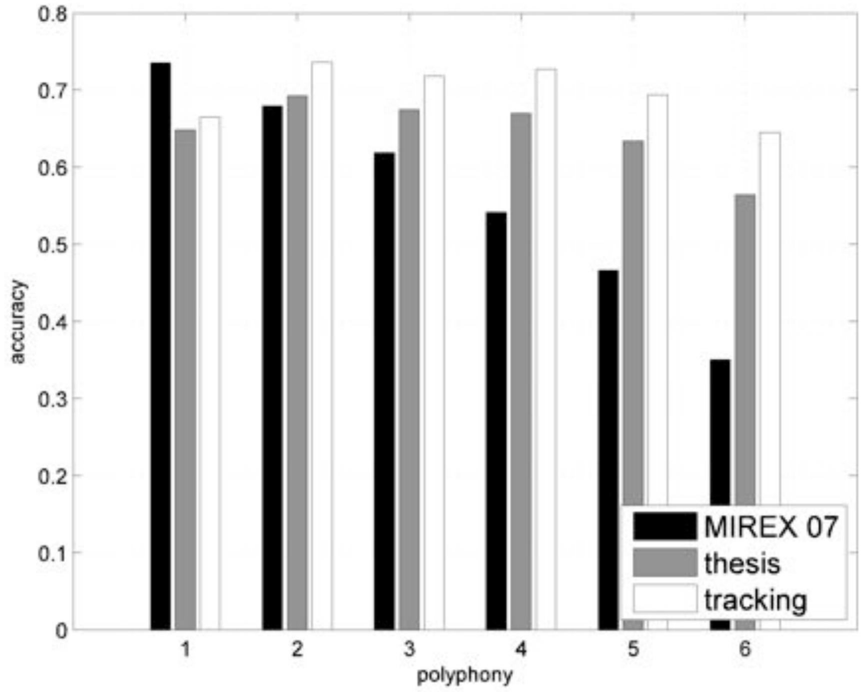


Evaluation

- Database:
  - Using RWC MIDI files and RWC Musical Instrument Sound Samples to synthesize polyphonic music [Yeh 07].
- Parameter set $(\lambda, \alpha, s, \mathrm{order})$
  - Trained by evolutionary algorithm
- Evaluation metrics

$$Acc = \frac{N_{corr}}{N_{corr} + N_{miss} + N_{subs} + N_{inst}}$$

Evaluation



- MIREX’07
  - 56.56%
- Yeh thesis ’08
  - 64.75%
- Tracking
  - 69.79%

Sound Example: Original 🔊 🔊 Transcribed 🔊 🔊

Conclusion

- Three possibilities to improve the system
  - Generate λ based on a transient or onset feature, instead of a fixed probability.
  - Consider octave streams to improve the iterative pruning process.
  - Share nodes in different paths to allow the intersection of source streams.

Conclusion

- Perspectives
  - The tracking architecture is generic and easy to implement
  - Mulitple-F0 estimation can be more efficient and robust by tracking F0 candidates beforehand