EE E6820: Speech & Audio Processing & Recognition

Lecture 7: Music analysis and synthesis

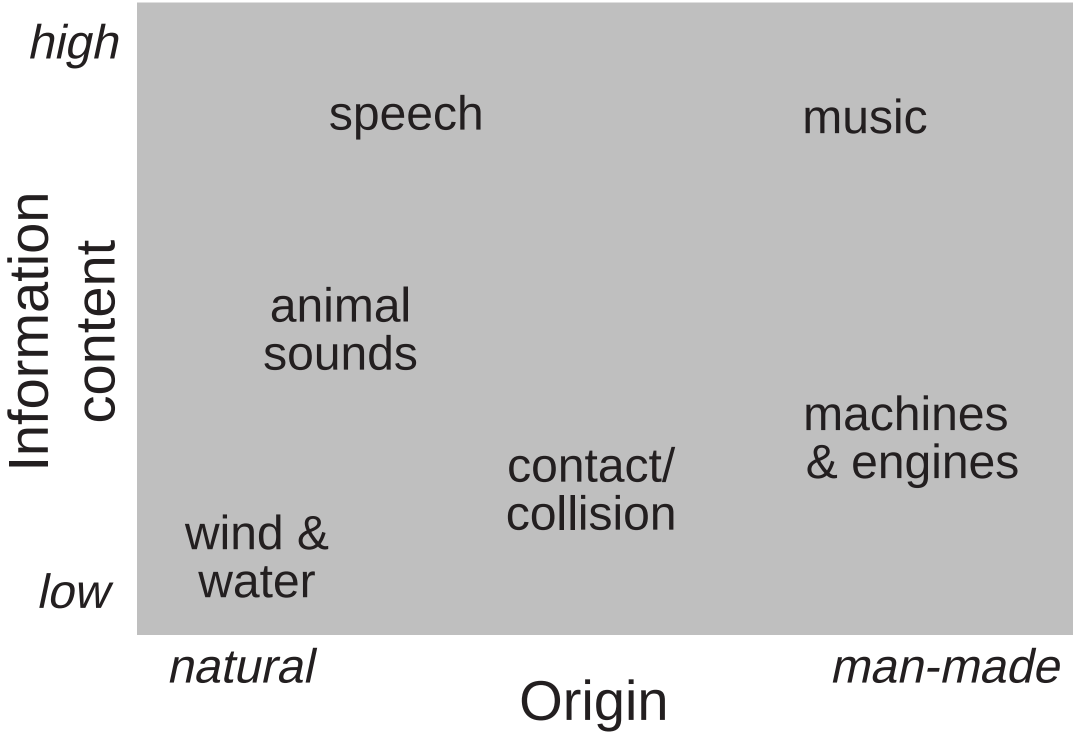
1. **Music and nonspeech**
2. **Music synthesis techniques**
3. **Sinewave synthesis**
4. **Music analysis**
5. **Transcription**

Dan Ellis <dpwe@ee.columbia.edu>
http://www.ee.columbia.edu/~dpwe/e6820/

1

Music & nonspeech

- **What is ‘nonspeech’?**
  - according to research effort: a little music
  - in the world: most everything



high

low

Information content

speech

music

animal sounds

machines & engines

contact/ collision

wind & water

natural

Origin

man-made

attributes?

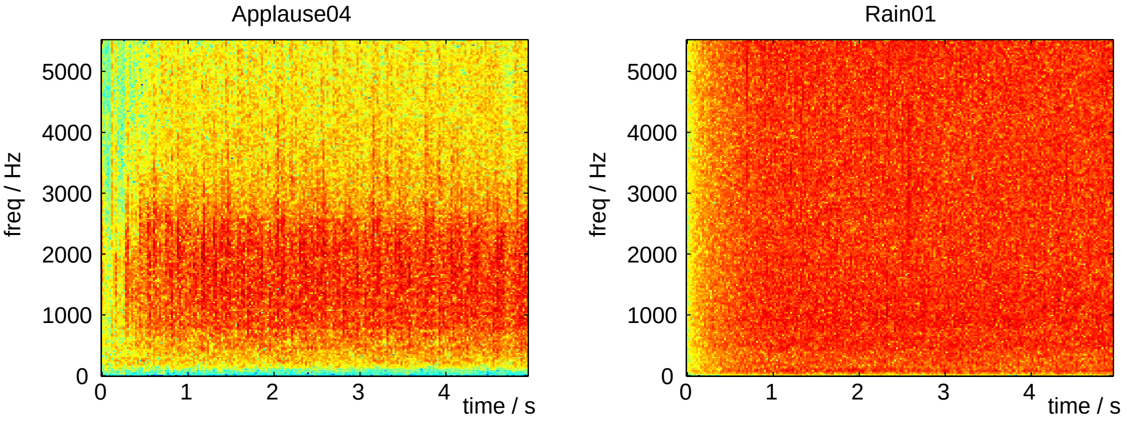
Sound attributes

- **Attributes suggest model parameters**
- **What do we notice about ‘general’ sound?**
  - psychophysics: pitch, loudness, ‘timbre’
  - bright/dull; sharp/soft; grating/soothing
  - sound is not ‘abstract’:
    tendency is to describe by source-events
- **Ecological perspective**
  - what matters about sound is ‘what happened’

  → our percepts express this more-or-less directly

Aside: Sound textures

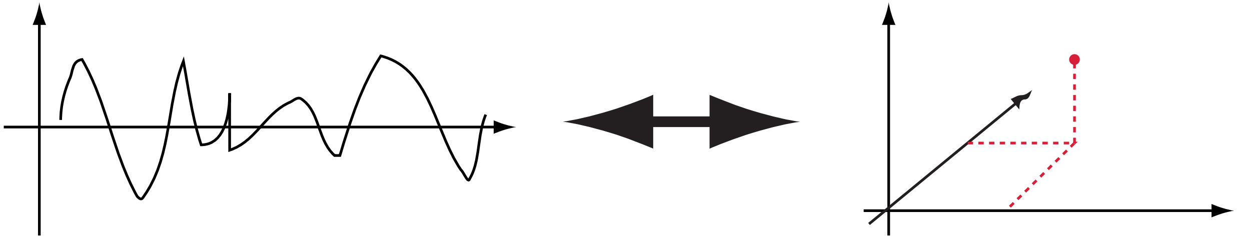
- **What do we hear in:**
  - a city street
  - a symphony orchestra
- **How do we distinguish:**
  - waterfall
  - rainfall
  - applause
  - static



- ***Levels* of ecological description...**

Motivations for modeling

- **Describe/classify**
  - cast sound into model because want to use the resulting parameters
- **Store/transmit**
  - model implicitly exploits limited structure of signal
- **Resynthesize/modify**
  - model separates out interesting parameters

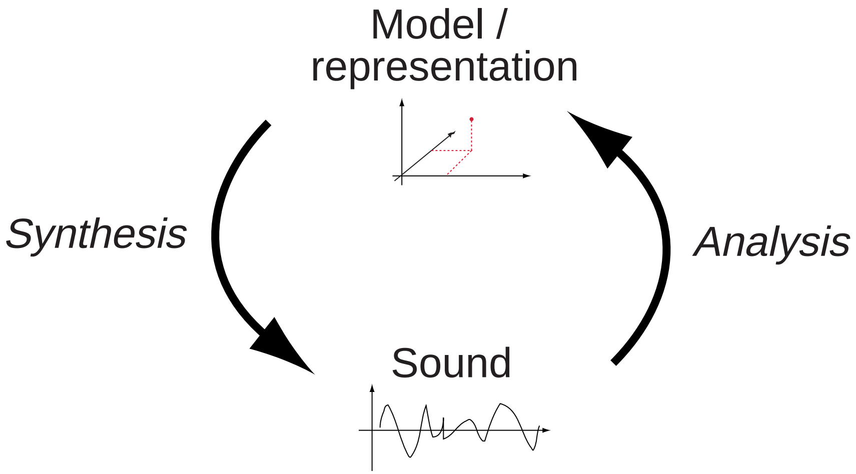


Sound

Model parameter space

Analysis and synthesis

- **Analysis is the converse of synthesis:**



Model / representation

Synthesis

Analysis

Sound

- **Can exist apart:**
  - analysis for classification
  - synthesis of artificial sounds

- **Often used together:**
  - encoding/decoding of compressed formats
  - resynthesis based on analyses
  - *analysis-by-synthesis*

Outline

1 Music and nonspeech

2 **Music synthesis techniques**
- Framework
- Historical development

3 Sinewave synthesis

4 Music analysis

5 Transcription

elements?

2

Music synthesis techniques

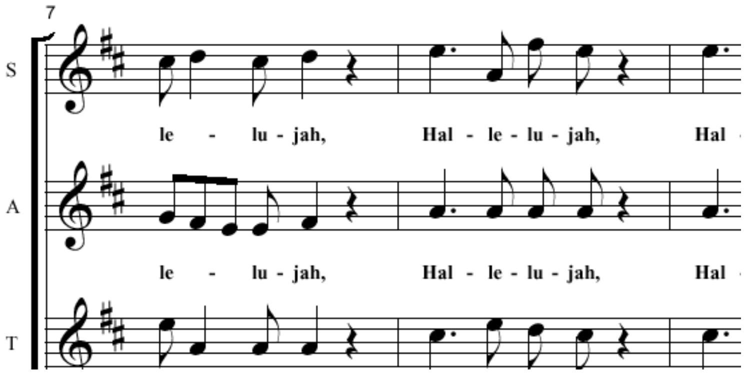
- **What is music?**
  - could be anything → flexible synthesis needed!
- **Key elements of conventional music**
  - instruments

  →note-events (time, pitch, accent level)

  →melody, harmony, rhythm
  - patterns of repetition & variation
- **Synthesis framework:**

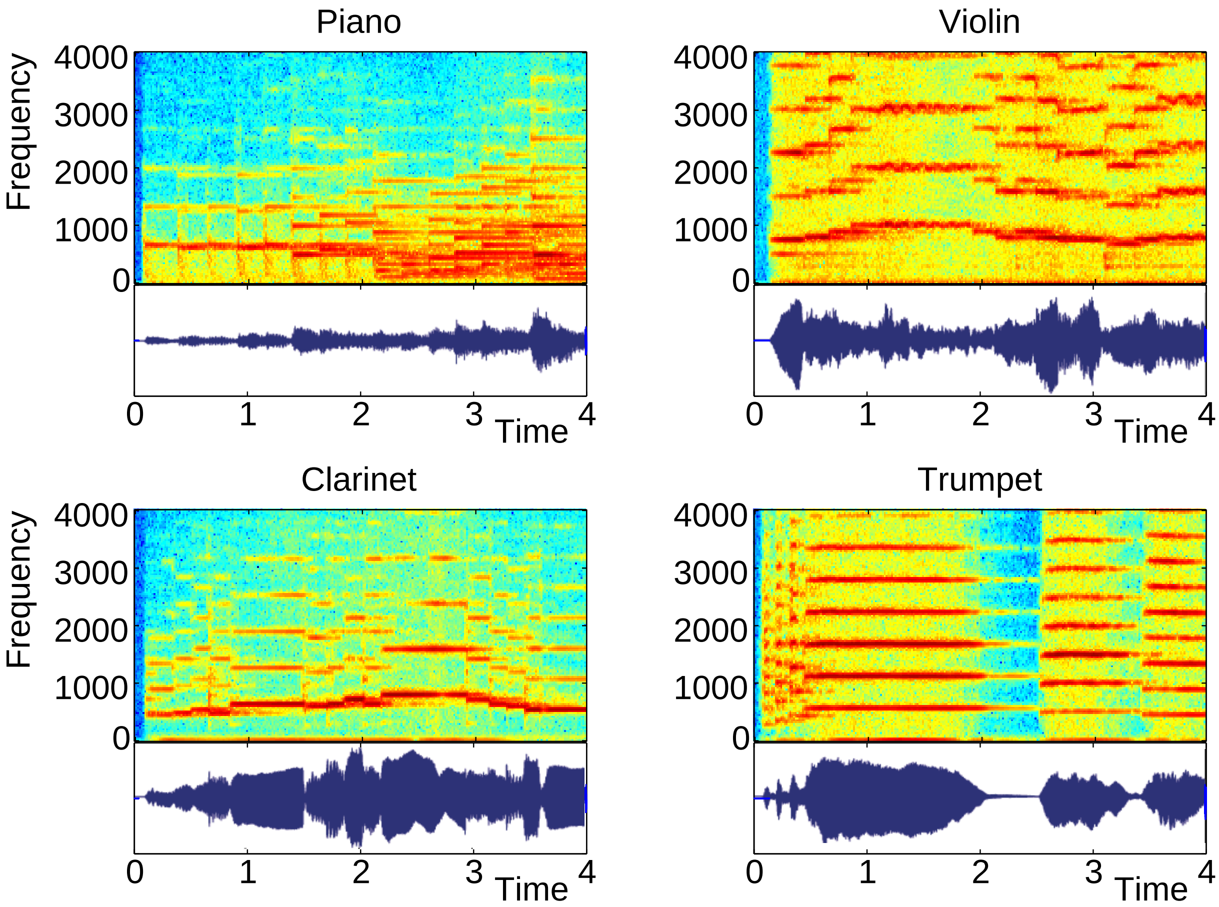
  instruments: common framework for many notes

  score: sequence of (time, pitch, level) note events


7
S
le - lu - jah, Hal - le - lu - jah, Hal
A
le - lu - jah, Hal - le - lu - jah, Hal
T

The nature of musical instrument notes

- **Characterized by instrument (register), note, loudness (emphasis), articulation...**



distinguish how?

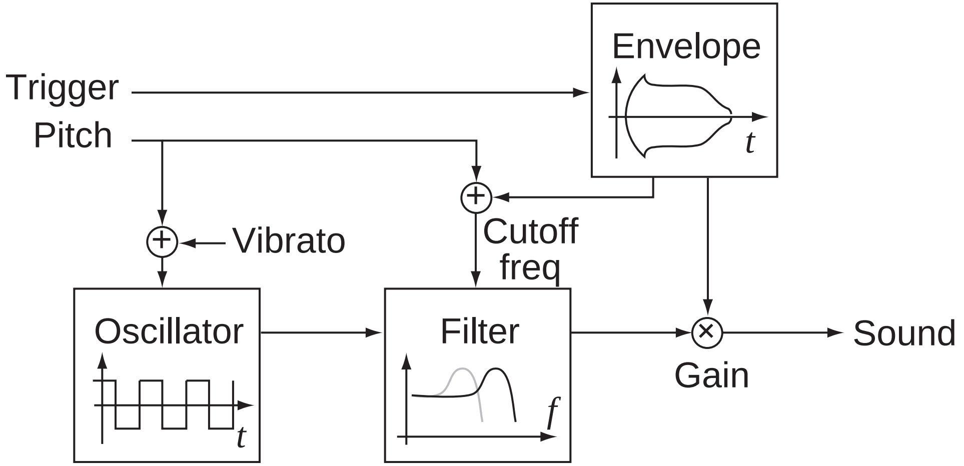
Development of music synthesis

- **Goals of music synthesis:**
  - generate realistic / pleasant new notes
  - control / explore timbre (quality)
- **Earliest computer systems in 1960s (voice synthesis, algorithmic)**
- **Pure synthesis approaches:**
  - 1970s: Analog synths
  - 1980s: FM (Stanford/Yamaha)
  - 1990s: Physical modeling, hybrids
- **Analysis-synthesis methods:**
  - sampling / wavetables
  - sinusoid modeling
  - harmonics + noise (+ transients)

others?

Analog synthesis

- **The minimum to make an ‘interesting’ sound**



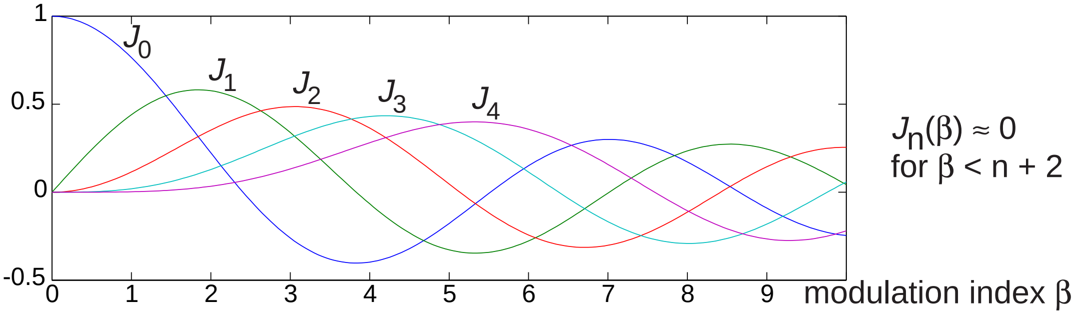
- **Elements:**
  - harmonics-rich oscillators
  - time-varying filters
  - time-varying envelope
  - modulation: low frequency + envelope-based
- **Result:**
  - time-varying spectrum, independent pitch

FM synthesis

- **Fast freq. modulation → harmonic sidebands:**

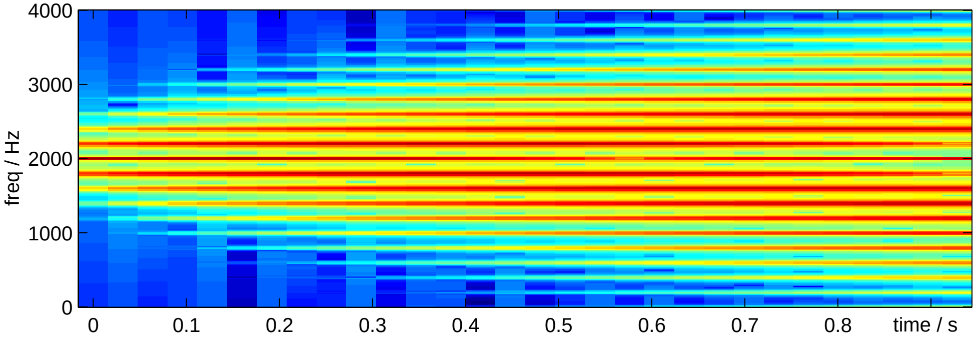
$$\cos(\omega_c t + \beta \sin \omega_m t) = \sum_{n=-\infty}^{\infty} J_n(\beta) \cos(\omega_0 + n\omega_m)$$

- **$J_n(\beta)$ is a Bessel function:**



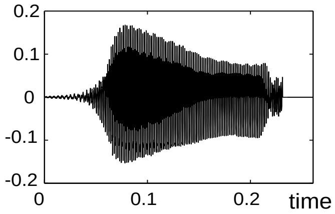
$J_n(\beta) \approx 0$
for $\beta < n + 2$

→ Complex harmonic spectra by varying β



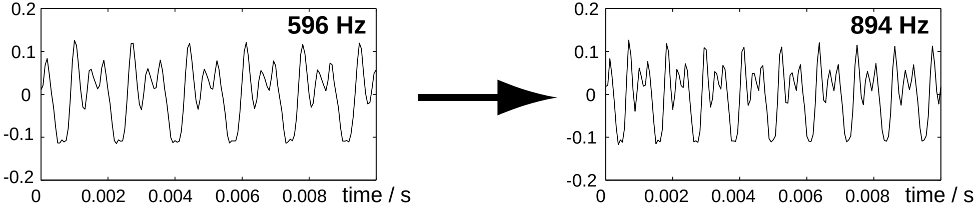
what use?

Sampling synthesis

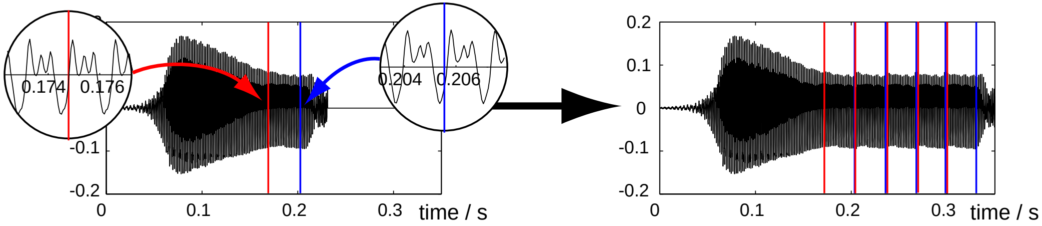


- **Resynthesis from real notes**
  → vary pitch, duration, level

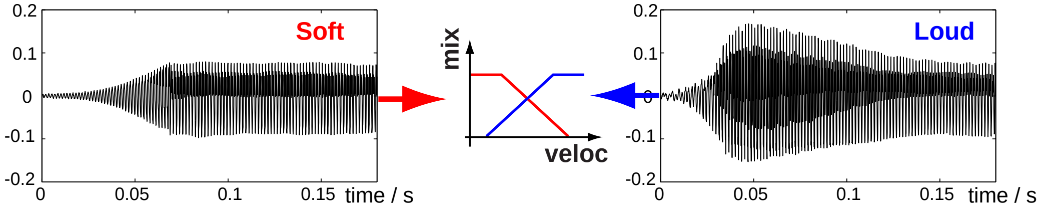
- **Pitch:** stretch (resample) waveform



- **Duration:** loop a ‘sustain’ section



- **Level:** cross-fade different examples



good
& bad?

  - need to ‘line up’ source samples

Outline

1. Music and nonspeech
2. Music synthesis techniques
3. **Sinewave synthesis** (detail)
   - Sinewave modeling
   - Sines + residual ...
4. Music analysis
5. Transcription

3 Sinewave synthesis

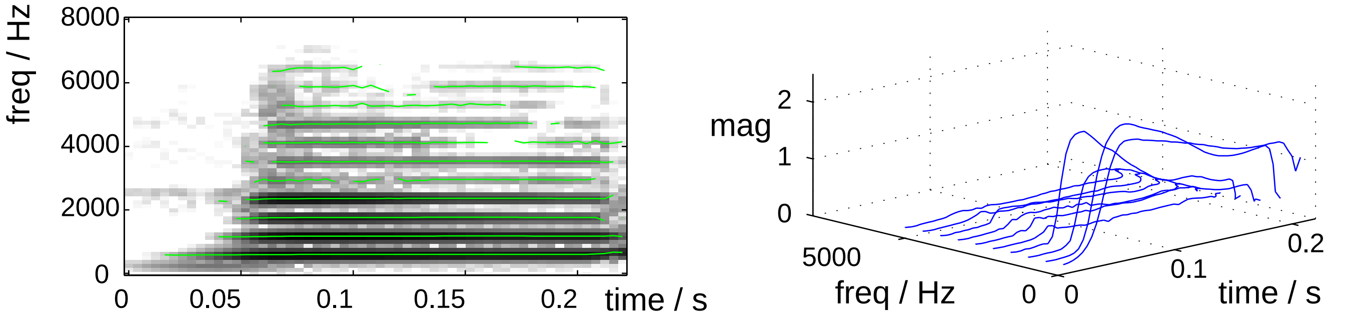
- **If patterns of harmonics are what matter, why not generate them all explicitly:**

  $$s[n] = \sum_k A_k[n] \cos(n \cdot k \cdot \omega_0)$$

  - particularly powerful model for pitched signals

- **Analysis (as with speech):**
  - find peaks in STFT $|S[\omega,n]|$ & track
  - or track fundamental ω_0 (harmonics / autoco) & sample STFT at k·ω_0

  →set of $A_k[n]$ to duplicate tone:



- **Synthesis via bank of oscillators**

Steps to sinewave modeling - 1

- **The underlying STFT:**

$$X[k, n_0] = \sum_{n=0}^{N-1} x[n + n_0] \cdot w[n] \cdot \exp -j\left(\frac{2\pi kn}{N}\right)$$

What value for N (FFT length & window size)?
What value for H (hop size: $n_0 = r \cdot H$, $r = 0, 1, 2\ldots$)?

- **STFT window length determines freq. resol'n:**

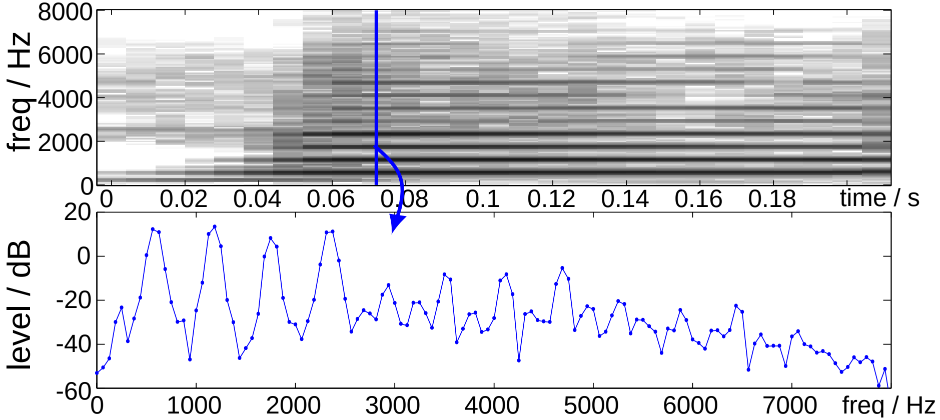
$$X_w(e^{j\omega}) = X(e^{j\omega}) * W(e^{j\omega})$$

- **Choose N long enough to resolve harmonics → 2-3x longest (lowest) fundamental period**
  - e.g. 30-60 ms = 480-960 samples @ 16 kHz
  - choose $H \leq N/2$

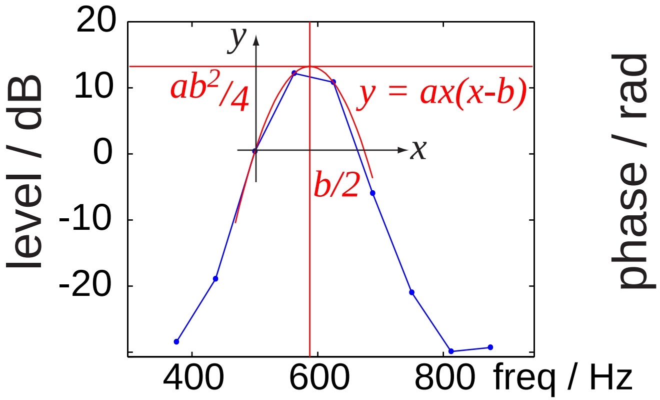
- **N too long → lost time resolution**
  - limits sinusoid amplitude rate of change

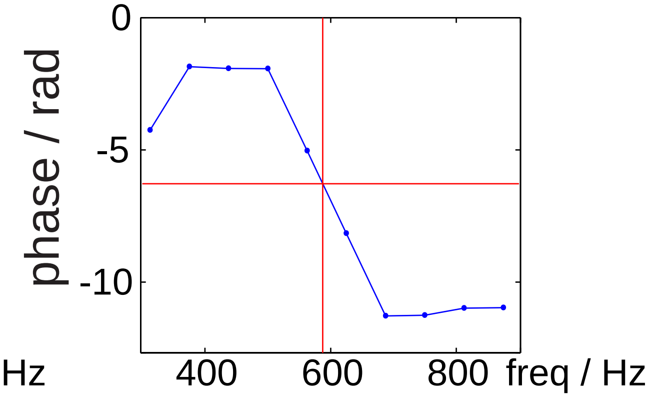
Steps to sinewave modeling - 2

- **Choose candidate sinusoids at each time by picking peaks in each STFT frame:**



- **Quadratic fit for peak, lin. interp. for phase:**

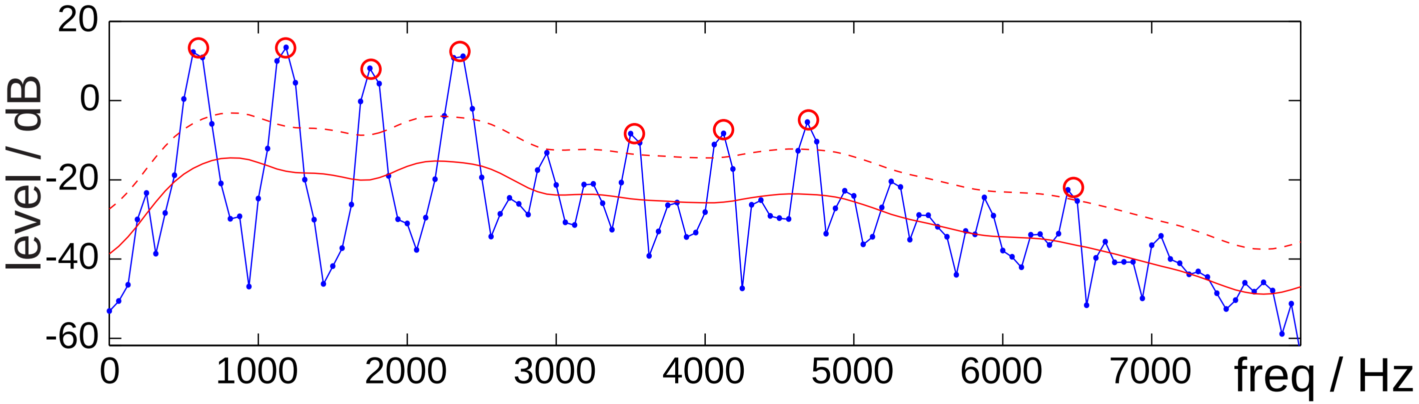




+ linear interp. of unwrapped phase

Steps to sinewave modeling - 3

- **Which peaks to pick?**
  Want ‘true’ sinusoids, not noise fluctuations
  - ‘prominence’ threshold above smoothed spec.

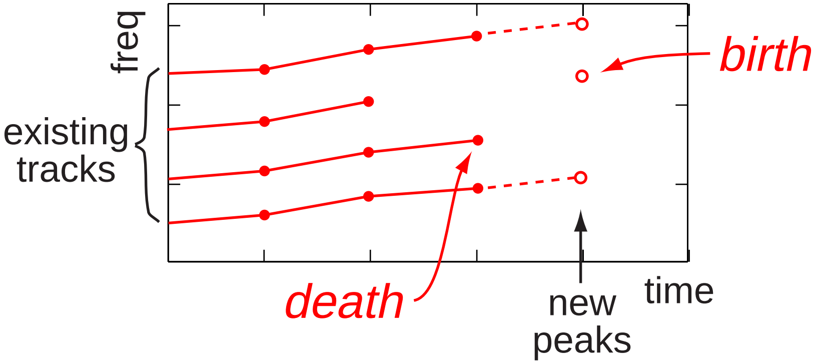


- **Sinusoids exhibit stability...**
  - of amplitude in time
  - of phase derivative in time

→compare with adjacent time frames to test?

Steps to sinewave modeling - 4

- **‘Grow’ tracks by appending newly-found peaks to existing tracks:**



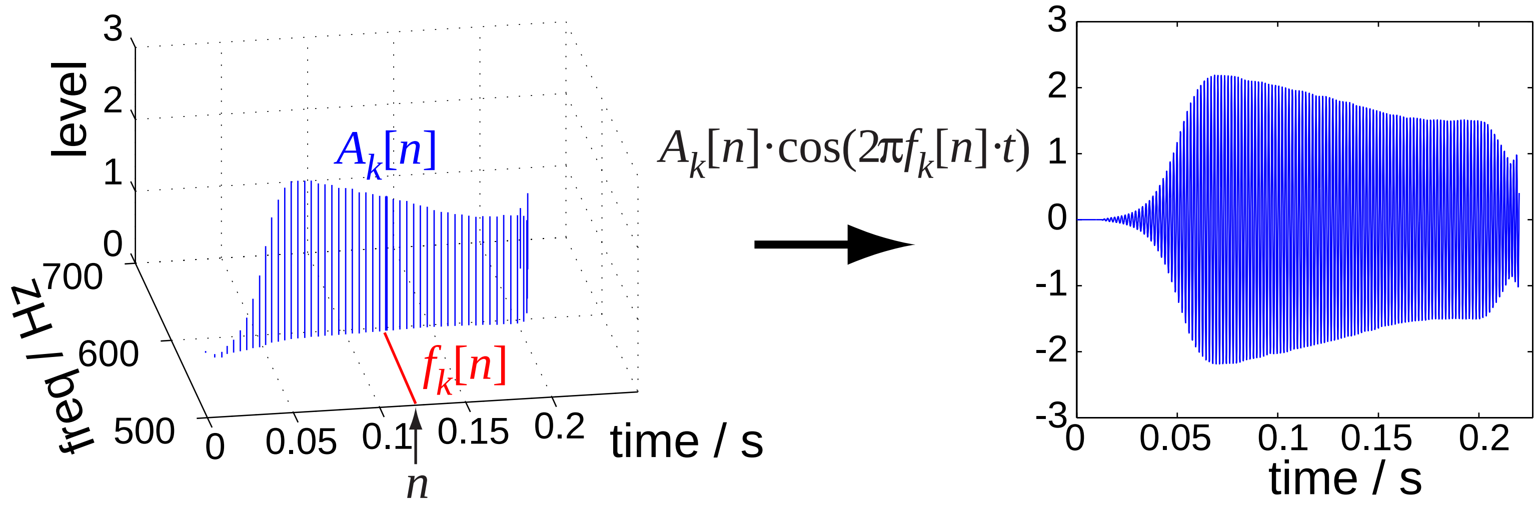
  - ambiguous assignments possible

- **Unclaimed new peak**
  - ‘birth’ of new track
  - backtrack to find earliest trace?

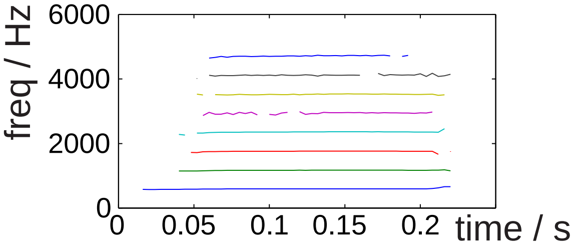
- **No continuation peak for existing track**
  - ‘death’ of track
  - or: reduce peak threshold for *hysteresis*

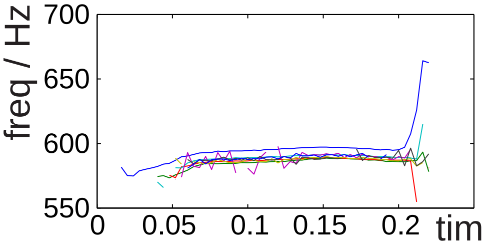
Resynthesis of sinewave models

- **After analysis, each track defines contours in frequency, amplitude $f_k[n]$, $A_k[n]$ (+ phase?)**
  - use to drive a sinewave oscillators & sum up



- **‘Regularize’ to exactly harmonic $f_k[n] = k\cdot f_0[n]$**

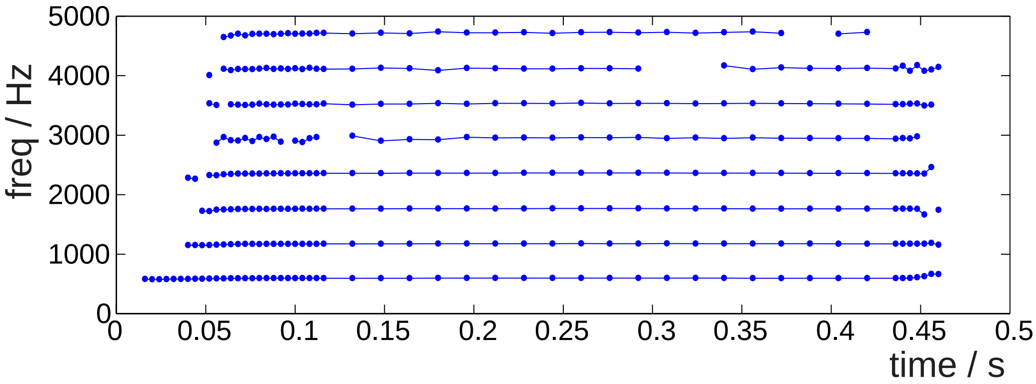




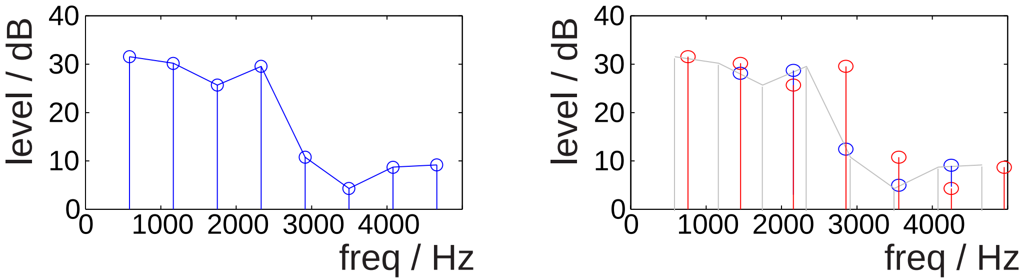
what
to do?

Modification in sinewave resynthesis

- **Change duration by warping timebase**
  - may want to keep onset unwarped



- **Change pitch by scaling frequencies**
  - either stretching or resampling envelope



- **Change timbre by interpolating params**

Sinusoids + residual

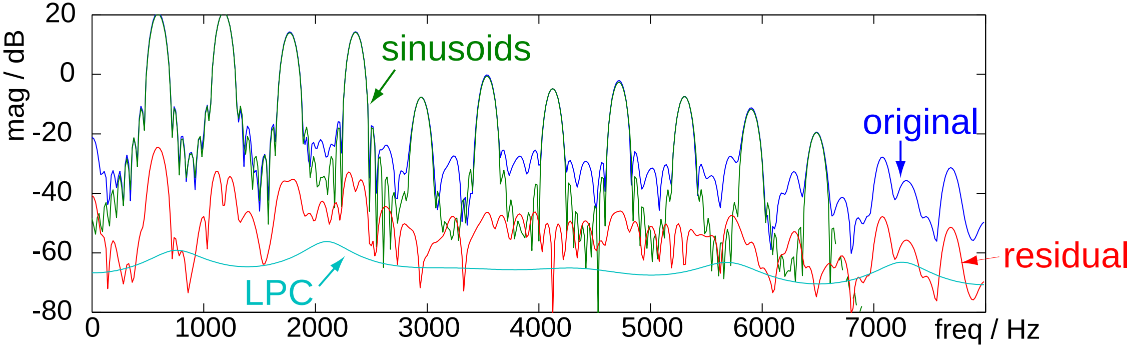
- **Only 'prominent peaks' became tracks**
  - remainder of spectral energy was noisy?

  → model residual energy with noise!

- **How to obtain 'non-harmonic' spectrum?**
  - zero-out spectrum near extracted peaks?
  - or: resynthesize (exactly) & subtract waveforms

$$e_s[n] = s[n] - \sum_k A_k[n] \cos(2\pi n \cdot f_k[n])$$

.. must preserve phase!



- **Can model residual signal with LPC**

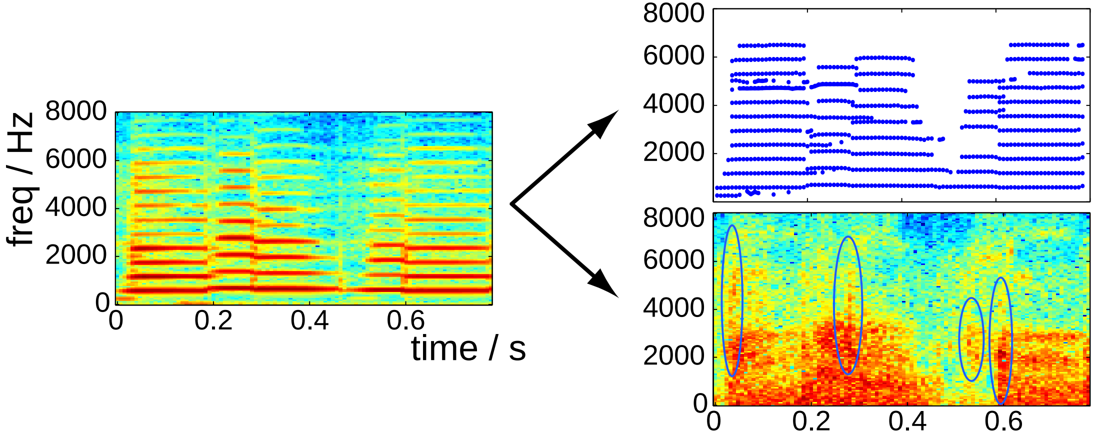
  →flexible representation of noisy residual

Sinusoids + noise + transients

- **Sound represented as sinusoids and noise:**

$$s[n] = \underbrace{\textstyle\sum_k A_k[n]\cos(2\pi n \cdot f_k[n])}_{\textbf{Sinusoids}} + \underbrace{h_n[n] * b[n]}_{\textbf{Residual } e_s[n]}$$

Parameters are $\{A_k[n], f_k[n]\}, h_n[n]$



$\{A_k[n], f_k[n]\}$

$h_n[n]$

- **Separate out abrupt transients in residual?**

$$e_s[n] = \textstyle\sum_k t_k[n] + h_n[n] * b[n]$$

  - more specific → more flexible

Outline

1. Music and nonspeech
2. Music synthesis techniques
3. Sinewave synthesis
4. **Music analysis**
    - Instrument identification
    - Pitch tracking
5. Transcription

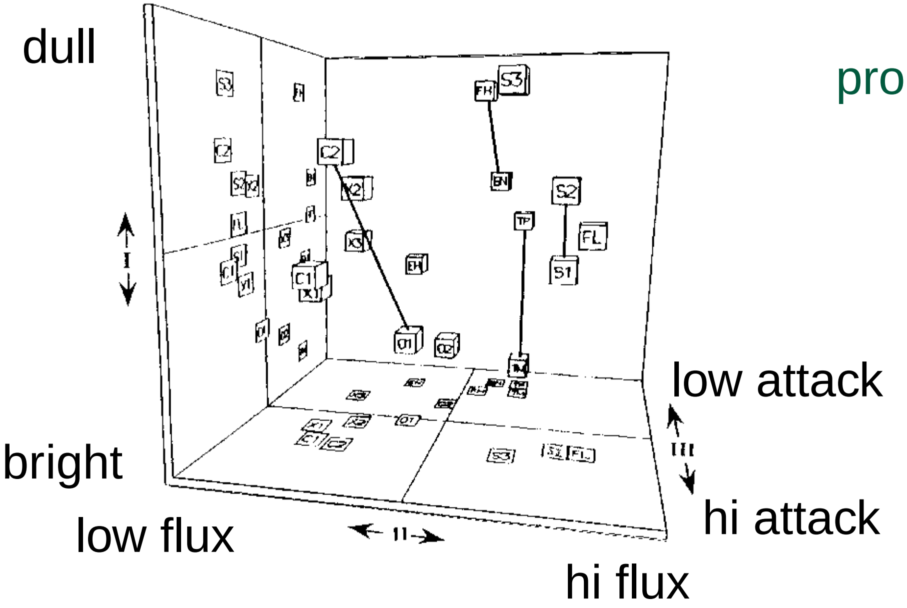
4

Music analysis

- **What might we want to get out of music?**
- **Instrument identification**
  - different levels of specificity
  - ‘registers’ within instruments
- **Score recovery**
  - transcribe the note sequence
  - extract the ‘performance’
- **Ensemble performance**
  - ‘gestalts’: chords, tone colors
- **Broader timescales**
  - phrasing & musical structure
  - artist / genre clustering and classification

Instrument identification

- **Research looks for perceptual ‘timbre space’**



procedure?

- **Cues to instrument identification**
  - onset (rise time), sustain (brightness)
- **Hierarchy of instrument families**
  - strings / reeds / brass
  - optimize features at each level

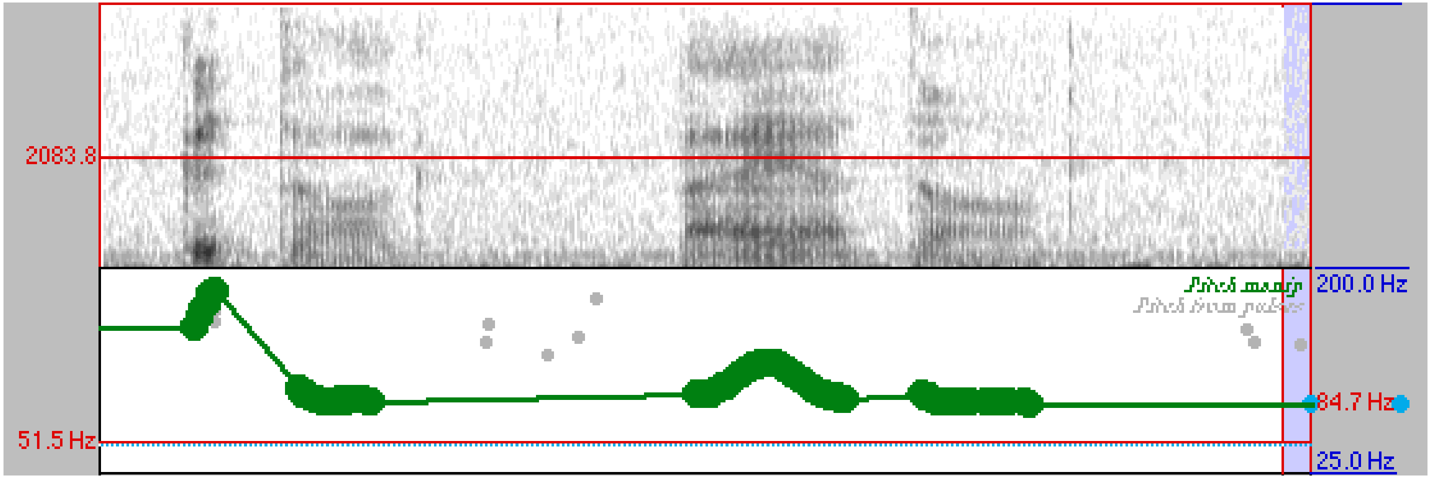
Pitch tracking

- **Fundamental frequency (→ pitch) is a key attribute of musical sounds**

  →pitch tracking as a key technology

- **Pitch tracking for speech**
  - voice pitch & spectrum highly dynamic
  - speech is voiced and unvoiced

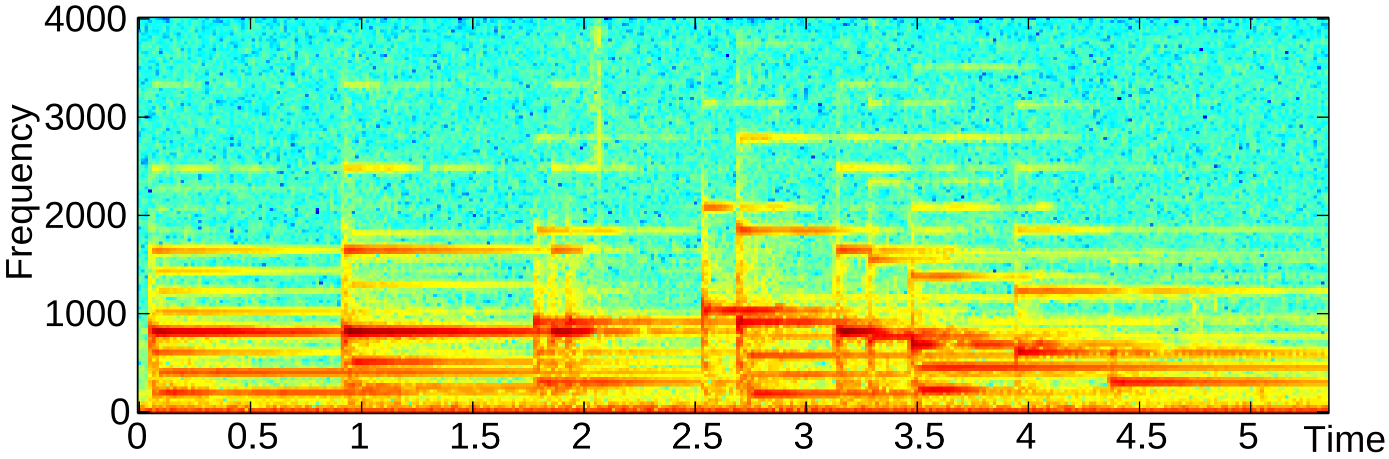
ground truth?



- **Applications**
  - voice coders (excitation description)
  - harmonic modeling

Pitch tracking for music

- **Pitch in music**
  - pitch is more stable (although vibrato)
  - but: *multiple pitches*



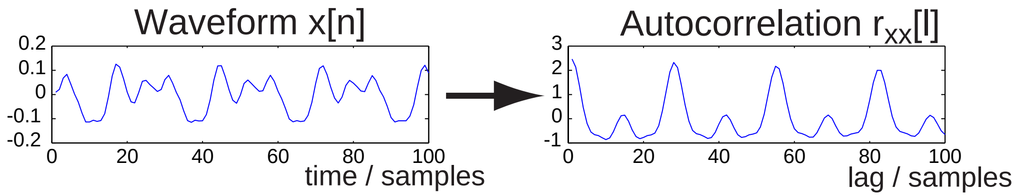
??

- **Applications**
  - harmonic modeling
  - music transcription (→ storage, resynthesis)
  - source separation
- **Approaches: “place” & “time”**

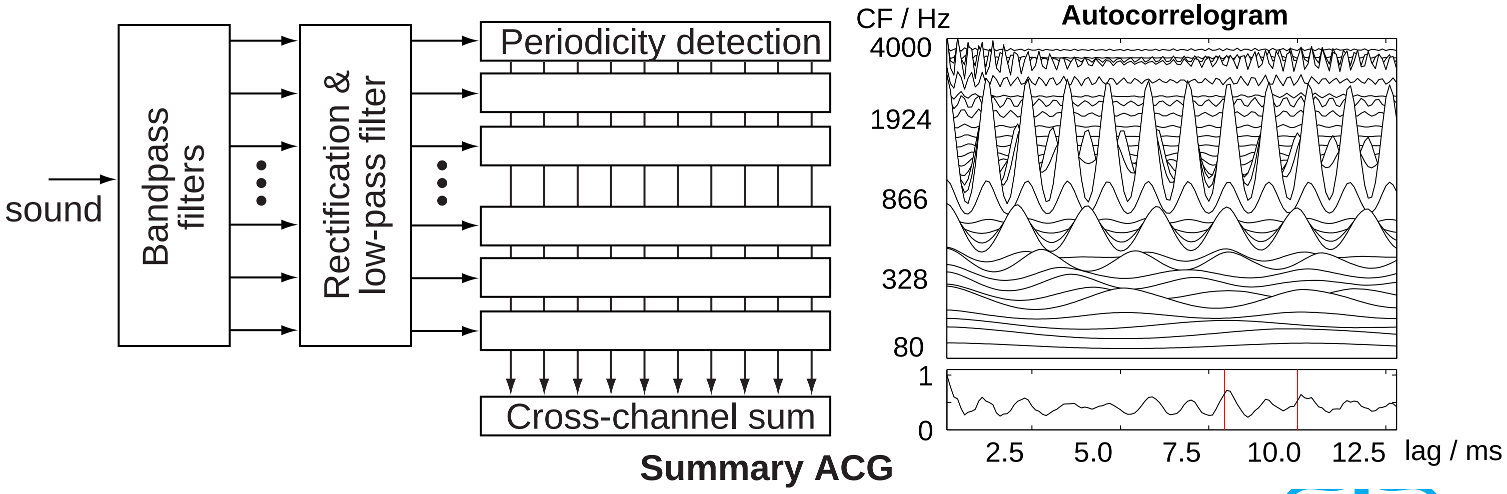
Meddis & Hewitt pitch model

- **Autocorrelation (time) based pitch extraction**
  - fundamental period → peak(s) in autocorrelation

$$x(t) \approx x(t+T) \quad \rightarrow \quad r_{xx}(T) = \int x(t)x(t+T) \approx max$$

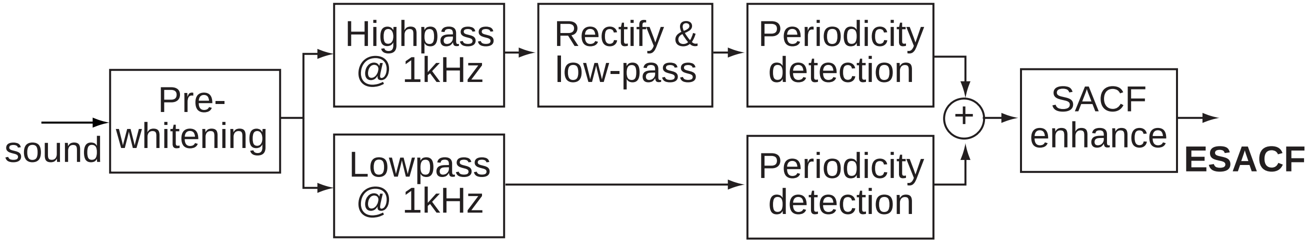


- **Compute separately in each frequency band & ‘summarize’ across (perceptual) channels**



Tolonen & Karjalainen simplification

- Multiple frequency channels can have different pitches dominant...
- But equalizing (flattening) the spectrum works:



sound

Pre-whitening

Highpass @ 1kHz

Rectify & low-pass

Periodicity detection

Lowpass @ 1kHz

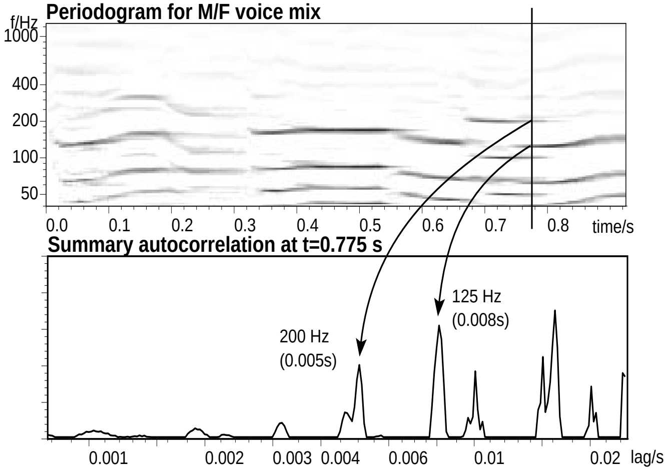
Periodicity detection

+

SACF enhance

ESACF

→ Summary AC as a function of time:

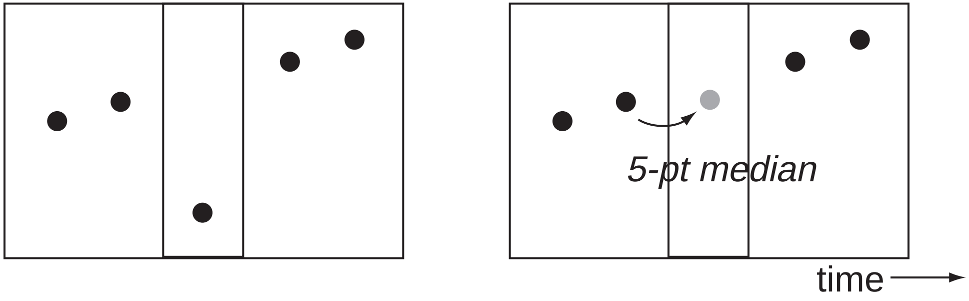


lag vs. freq?

- ‘Enhancement’ = cancel *subharmonics*

Post-processing of pitch tracks

- **Remove outliers with median filtering**



- **Octave errors are common:**
  - if $x(t) \approx x(t + T)$ then $x(t) \approx x(t + 2T)$ etc.

  → **dynamic programming/HMM**

- **Validity**
  - “is there a pitch at this time?”
  - voiced/unvoiced decision for speech

- **Event detection**
  - when does a pitch slide indicate a new note?

1 Music and nonspeech

2 Music synthesis techniques

3 Sinewave synthesis

4 Music analysis

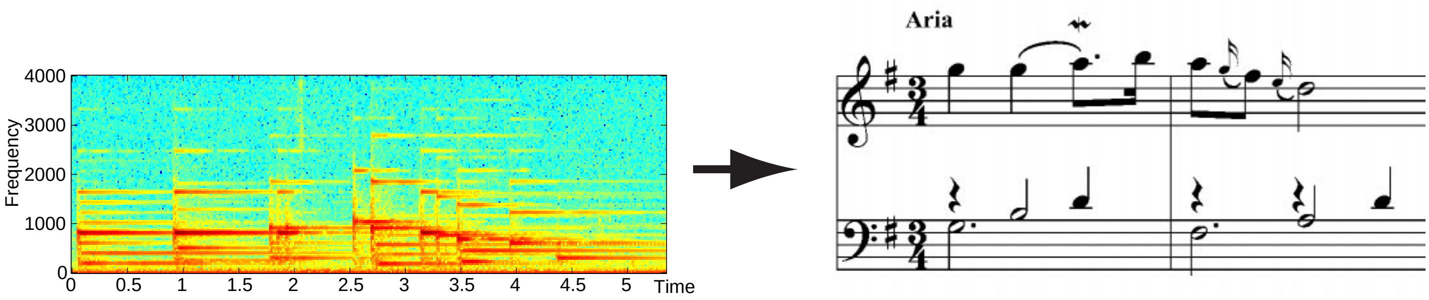
5 Transcription

- Bottom-up and top-down
- Transcription from sinewave models

5

Transcription

- **Basic idea: Recover the score**



- **Is it possible? Why is it hard?**
  - music students do it
    ... but they are highly trained; know the rules

- **Motivations**
  - for study: what was played?
  - highly compressed representation (e.g. MIDI)
  - the ultimate restoration system...

Transcription framework

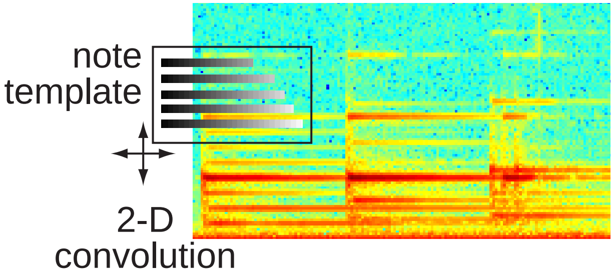
- **Recover discrete events to explain signal**

Note events $\xrightarrow{\textit{synthesis}}$? Observations

$\{t_k, p_k, i_k\}$ $X[k,n]$

  - analysis-by-synthesis?

- **Exhaustive search?**
  - would be possible given *exact note waveforms*
  - .. or just a 2-dimensional ‘note’ template?



but superposition is not linear in $|STFT|$ space

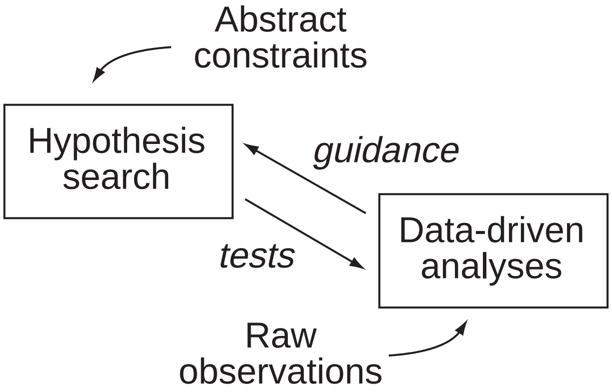
- **Inference depends on all detected notes**
  - is this evidence ‘available’ or ‘used’?
  - full solution is exponentially complex

Bottom-up versus top-down

- **Bottom-up: observ’n directly gives description**
  - e.g. peaks in 2-D convolution
  - but: few domains are that ‘linear’

- **Top-down: pursue & confirm *hypotheses***
  - e.g. analysis-by-resynthesis matching
  - but: need to limit search space

- **Generally, need to do both:**

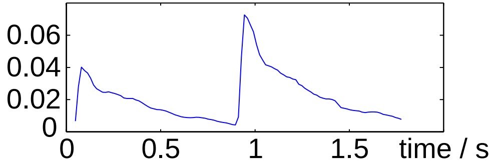


  - bottom-up guides & limits search
  - top-down resolves ambiguities in low-level

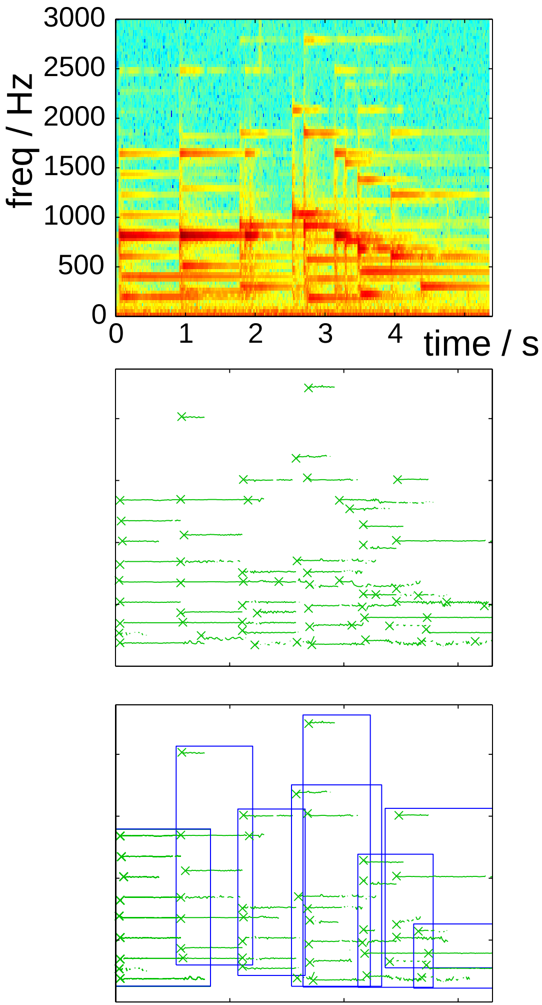
how to transcribe?

Transcription from sinewave models

- **Form sinusoid model**
  - as with synthesis, but signal is more complex
- **Break tracks**
  - need to detect new ‘onset’ at single frequencies



- **Group by onset & common harmonicity**
  - find sets of tracks that start around the same time
  - + stable harmonic pattern
- **Pass on to constraint-based filtering...**



bu/td? mistakes?

Problems for transcription

- **Music is practically worst case!**
  - note events are often synchronized
    → defeats common onset
  - notes have harmonic relations (2:3 etc.)
    → collision/interference between harmonics
  - variety of instruments, techniques, ...
- **Listeners are very sensitive to certain errors**
  - .. and impervious to others
- **Apply further constraints**
  - like our ‘music student’
  - maybe even the whole score (Scheirer)!

Summary

- **‘Nonspeech audio’**
  - i.e. sound in general
  - characteristics: ecological
- **Music synthesis**
  - control of pitch, duration, loudness, articulation
  - evolution of techniques
  - sinusoids + noise + transients
- **Music analysis**
  - different aspects: instruments, pitches, performance
  - transcription complications: representation, octaves, onsets, ...
  - rely on high-level structural constraints

and beyond?