Real-time pitch tracking

Contents

- Pitch definitions
- Applications
- Real-time method requirements
- Algorithm examples
  - Time-domain methods
  - Frequency-domain methods
  - Statistical methods
- General improvements
- Method evaluation

Definitions

- Instant frequency ω_i in the case of pseudo-periodic sounds

$$x(t) = \sum_{i=1}^{I(t)} A_i(t) \exp(j\phi_i(t)) \quad \text{with} \quad \phi_i(t) = \int_{-\infty}^{t} \omega_i(\tau) d\tau$$

- Instant fundamental frequency
  - Shortest ω_i
- Modern pitch perception models:
  - Periodicity of neural patterns in the time domain (Licklider 1951)
  - Harmonic pattern of partials resolved by the cochlea in the frequency domain (Goldstein 1973)
- Other F*0* definitions:
  - Rate of vibrations of the vocal folds
  - Normalized definition
- Multiple pitch extraction

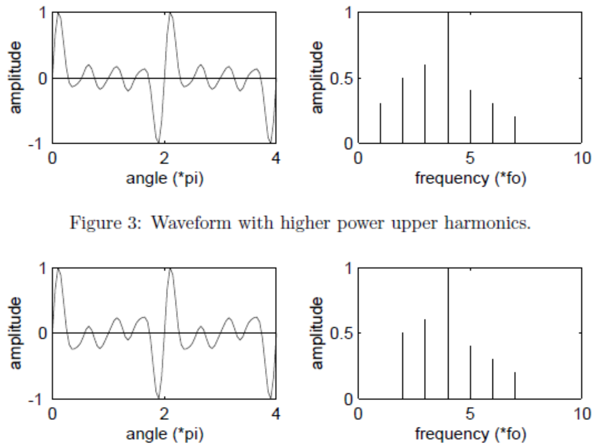


Figure 3: Waveform with higher power upper harmonics.

(Gerhard 2001)

Applications

- Original problems in speech processing:
  - Classification voiced/unvoiced signals
  - Speaker identification

- Music applications
  - Real-time music transcription
  - Audio-to-MIDI conversion
  - Pitch modification
    - PSOLA – Pitch Synchronous Overlap Add Method (Moulines and Charpentier 1990)
    - Lent’s algorithm (Lent 1989)

Realtime pitch tracking (Cuadra 2001)

- Problem solved for recorded monophonic voices or sounds

- Still difficult in live conditions

- Requirements:
  - Real-time functioning
  - Minimal output delay (latency)
  - Robustness (noise)
  - Sensitivity to musical requirements of the performance

Live pitch tracking requirements (Cuadra 2001)

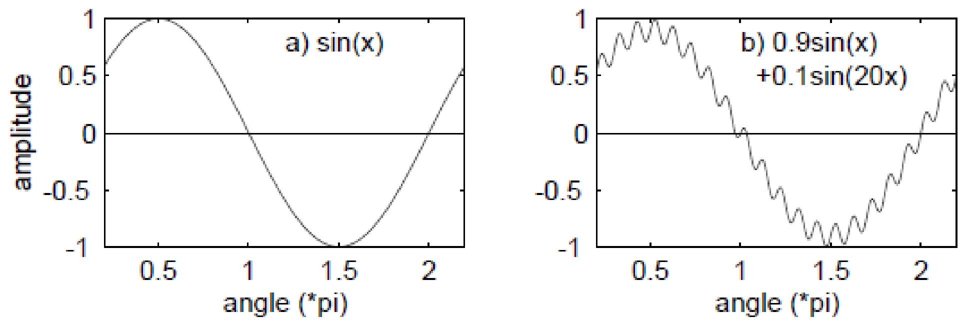
- Real-time functioning:
  - Error checking computational cost
  - Heavy overlapping of the frequency transforms
  - Several algorithms run in parallel
- Minimal output delay (latency)
  - Pitch-to-MIDI implementation
- Robustness (noise)
  - Performance environment
  - Recording equipment
- Sensitivity to musical requirements of the performance
  - frequency resolution of at least semi-tones, including the correct octave
  - timely recognition and quality of instantaneous pitch for possible real-time conversion into symbolic pitch
  - instruments with well-behaved harmonics (such as cello and flute).

Approaches

- Time domain
  - Zero-crossing rate analysis
  - Autocorrelation function
  - Instantaneous frequency detection
- Frequency domain
  - Harmonic period spectrum
  - Cepstrum analysis
  - Maximum likelihood
- Statistical
  - Neural networks
  - Hidden Markov Models

Zero-crossing rate (Gerhard 2001)

- Extracts the distance between two zero crossing as being the period related to the fundamental frequency
- Perform badly on inharmonic sounds or sounds with power in the higher frequencies
- Intrinsic information to be used with other algorithms



Weighted Autocorrelation Function (Kobayashi 1995; Cuadra 2001)

- Algorithm
  - pick peaks in the autocorrelation function…
  - …or in the average magnitude difference function…
  - …or with an improved estimator

$$\phi(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n+\tau)$$

$$\psi(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} |x(n) - x(n+\tau)|$$

$$f(\tau) = \frac{\phi(\tau)}{\psi(\tau) + k}$$

- Advantages
  - The last estimator is noise-robust
  - Efficient in the case of allowed gross pitch error (10 Hz)

Autocorrelation function - Algorithm (de Cheveigné and Kawahara 2001)

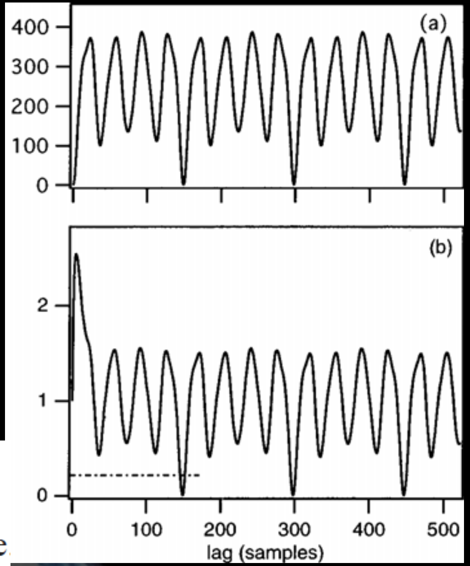
- Autocorrelation function
  - Octave errors

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau}$$

- Difference function

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2$$

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau)$$



- Cumulative mean normalized difference function
  - ➔ Less "too high" errors

$$d'_t(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_t(\tau) \Big/ \left[(1/\tau) \sum_{j=1}^{\tau} d_t(j) \right] & \text{otherwise} \end{cases}$$

- Absolute threshold for *d'*
  - ➔ Less "too low" errors
- Parabolic interpolation on *d*
  - ➔ Improve detection resolution
- Best local estimate of *d'*

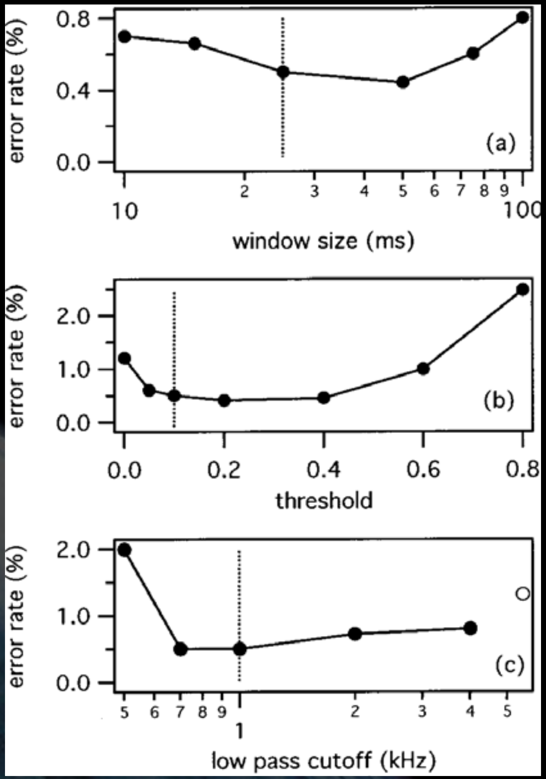
Version	Gross error (%)
Step 1	10,0
Step 2	1,95
Step 3	1,69
Step 4	0,78
Step 5	0,77
Step 6	0,50

Autocorrelation function (de Cheveigné and Kawahara 2001)

- Works well up to ¼ of the sampling frequency

- No need of detection upper limit

- Sensible to the definition of parameters



Pitch extraction based on instantaneous frequency (Abe et al. 1995)

- Band-pass filter bank
- Each of the filter is controlled to be tracking one harmonic component

$$\phi_n(t) = \frac{d}{dt}\arg[y_n(t)]$$

- The lowest frequency of each harmony determines the detected pitch

speaker	M1	M2	M3
proposed	0.170	0.204	0.164
cepstrum	0.284	0.300	0.320
speaker	F1	F2	F3
proposed	0.061	0.062	0.094
cepstrum	0.110	0.122	0.167

- No double-pitch or half-pitch errors

speaker	M1	M2	M3
proposed	0.000	0.000	0.000
cepstrum	0.845	0.868	1.257
speaker	F1	F2	F3
proposed	0.000	0.000	0.000
cepstrum	4.835	5.945	8.190

- Improvement by deducing the pitch from the harmonic spectrum (more robust)

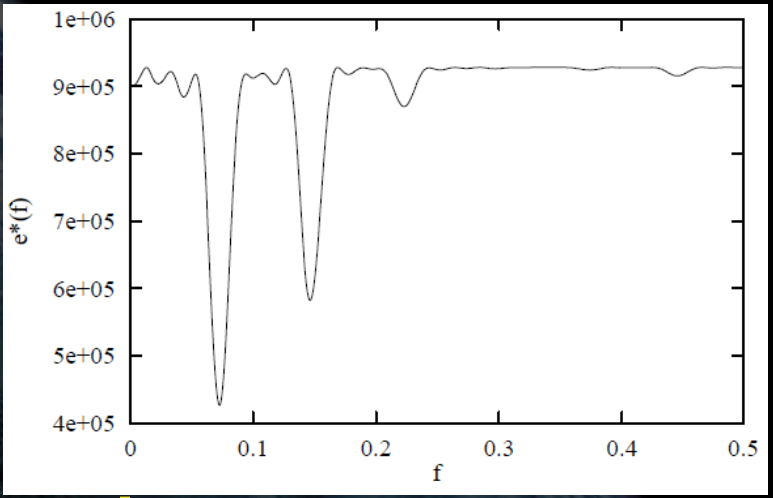
Pitch extraction by least-square fitting (Choi 1995)

- Evaluates the square error between the signal and a sinusoidal function

$$\hat{w}_k = a\sin(fk) + b\cos(fk)$$

$$e = \sum_{k=1}^{N} (\hat{w}_k - w_k)^2$$

- The estimate coefficients show peaks on signal harmonics



- The peak width allows to perform estimation on few frequencies
- The frequency is then extracted by interpolation
- No windowing is required

Harmonic Product Spectrum (Noll 1969; Cuadra 2001)

- Algorithm:
  - Measure the maximum coincidence for harmonics

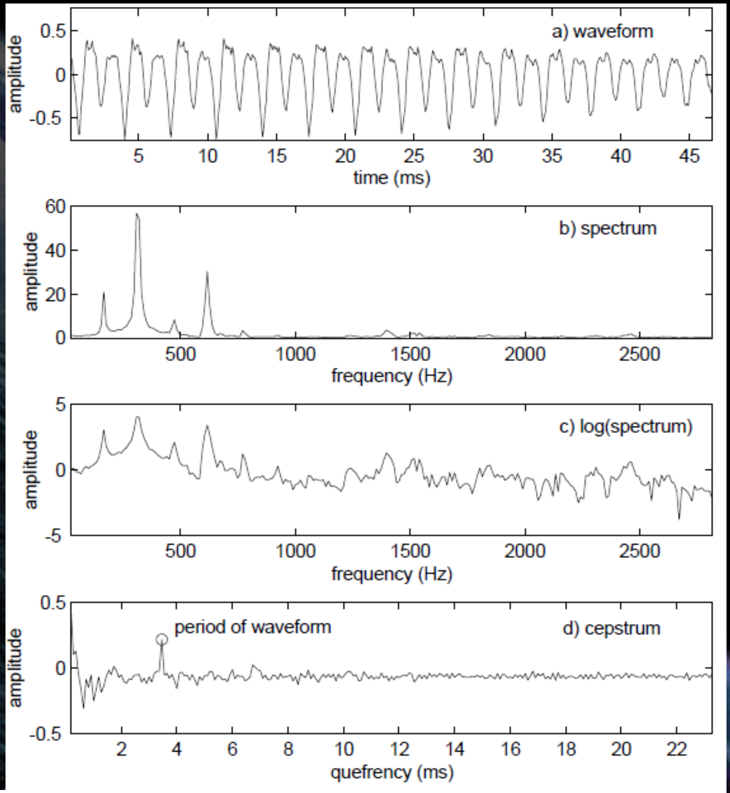
$$Y(\omega) = \prod_{r=1}^{R} |X(\omega r)|$$

$$\hat{Y} = \max_{\omega_i} \{Y(\omega_i)\}$$

- Advantages:
  - Works well under a wide range of conditions
- Drawbacks:
  - Need to enhance low frequency resolution with zero padding
  - Octave errors (generally one octave too high) ➔ post-processing
  - Errors for frequencies below 50 Hz due to noise

Cepstrum analysis (Noll 1967; Gerhard 2001)

- Algorithm
  - Cepstrum: signal synthesized from the log-magnitude of the signal Fourier transform
  - Search through the cepstrum a peak in a limited range, corresponding to the period of the signal
- Advantages:
  - Quite robust to noise
- Drawbacks:
  - Errors in the case of inharmonic sounds



(Gerhard 2001)

Maximum Likelihood (Noll 1969; Cuadra 2001)

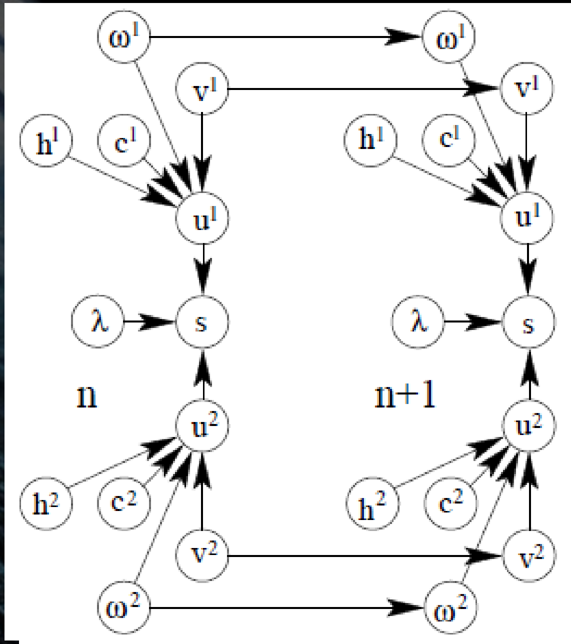
- Algorithm:
  - Search the best match through a set of possible ideal spectra

$$E(\omega) = \|Y - \tilde{Y_\omega}\|^2 \qquad \hat{Y} = \min_{\omega} \{E(\omega)\}$$

- Advantages:
  - No spectral interpolation needed ➔ smaller transform sizes
  - Works well up to one octave outside its range
- Drawbacks:
  - Efficiency of the algorithm ⇔ pitch resolution
  - Works well only with a fixed tuning (keyboards, woodwinds,…)
  - Less robust to noise and weak signals than the previous method

Statistical algorithms

- Use of the intrinsic temporal/frequency similarity between sounds of same pitch ➔ classification problem
- Requires an adapted training
- Neural networks for voiced/unvoiced classification (Barnard et al. 1991)
- Hidden Markov Models for one-singer and multi-singer pitch tracking (Bach and Jordan 2005)



(Bach and Jordan 2005)

General improvements

- Improvements can be added to lower the error rate of this algorithms
- Pre-processing (e.g., low-pass filtering)
- Post-processing (e.g., parabolic interpolation, pitch smoothing)
- Extra information (e.g., zero-crossing rate, auditory model)

Pitch extraction based on pitch perception model (de Cheveigné 1991)

- Use the average magnitude difference function

$$\psi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} |x(n) - x(n+\tau)|$$

- Based on the Licklider's perception model (Licklider 1951)
  - Apply a filter bank to the signal
  - Perform the autocorrelation test on each bands

- Quite weak efficiency

- Could be added as extra information in another algorithm

Algorithm evaluation

- Common errors:
  - Harmonic errors
  - Subharmonic errors
  - Transient signals
- Evaluation problem
  - Ground truth?
  - Consistency between estimators
  - Common database (Plante 1995)
- Comparison criteria
  - Gross error rate
  - Fine error rate
  - Difference between estimators

Conclusion

THANK YOU

References

Abe, T., T. Kobayashi, and S. Imai. 1995 Harmonics tracking and pitch extraction based on instantaneous frequency. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*: 756–59.

Bach, F., and M. Jordan. 2005. Discriminative training of Hidden Markov Models for multiple pitch tracking. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.

Barnard, E., R.A. Cole, M.P. Vea, and F.A. Alleva. 1991. Pitch detection with a neural-net classifier. *IEEE Transactions on Signal Processing* 39 (2): 298–307.

Choi, A. 1997. Real-time fundamental frequency estimation by least-square fitting. *IEEE Transactions on Speech and Audio Processing*: 201–5.

Cuadra, P., A. Master, and C. Sapp. 2001. Efficient pitch detection techniques for interactive music. *Proceedings of the International Computer Music Conference*.

de Cheveigné, A. 1991. Speech f0 extraction based on Licklider's pitch perception model. *Proceedings of the International Congresses of Phonetic Sciences*.

de Cheveigné, A., and H. Kawahare. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111 (4): 1917–30.

Gerhard, D. 2003. Pitch extraction and fundamental frequency: history and current techniques. Technical Report TR-CS 2003-6, University of Regina Department of Computer Science.

Goldstein, J. 1973. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustic Society of America* 54 (6): 1496–1516.

Kobayashi, H., and T. Shimamura. 1995. A weighted autocorrelation method for pitch extraction of noisy speech. *Proceedings of the Acoustical Society of Japan*: 343–4.

Lent, K. 1989. An efficient method for pitch shifting digitally sampled sounds. *Computer Music Journal* 13 (4).

Licklider, J. 1951. A duplex theory of pitch perception. *Experientia* 7 (4): 128–134.

Moulines, Eric, and Francis Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communications* 9 (5-6): 453–67.

Noll, M. 1967. Cepstrum pitch determination. *The Journal of the Acoustical Society of America* 41 (2): 293–309.

Noll, M. 1969. Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. *Proceedings of the Symposium on Computer Processing ing Communications*: 779–97.

Plante, F., G. Meyer, and W. Ainsworth. 1995. A pitch extraction reference database. *Proceedings of EUROSPEECH*.

Wise, J., J. Caprio, and T. Parks. 1976. Maximum likelihood pitch estimation. *IEEE Transaction on Acoustics, Speech, Signal Processing* 24 (5): 418–23.