Master M2 - DataScience

Audio and music information retrieval

Lecture on
Pitch and Multipitch estimation

Gaël RICHARD

Télécom Paris

March 2023

Content

- **Foreword on pitch perception**
  - Definition
  - Perception of tonal height of pure sounds
  - Pitch of complex sounds
- **Fundamental frequency detection**
- **Multipitch detection**
- **Machine learning based estimation methods**

Lecture 9: What you need to know

Pitch perception

- What is the difference between global and analytic pitch perception ?
- Explain pitch perception principles when some harmonics are missing

Pitch estimation

- Explain what is an autocorrelation-based method for pitch estimation
- Explain what is the spectral sum. Why is it appropriate for pitch estimation ?
- What is the output of a typical deep neural network (CNN based) for pitch estimation (output of last layer, dimension,..) ?

Multipitch estimation

- What is an iterative method for multi-pitch estimation ?
- What are the main difficulties of multipitch estimation ?

What is pitch / musical height perception ?

- **Zwicker sees 4 "heights":**
  - (physics) frequency (Hz)

    log -> harmonic height
  - (subjective) height / tonie

    log -> melodic height

- **But the « subjective height » (or pitch) covers 4 aspects:**
  - « Global » height
  - Tonal pitch
  - Spectral pitch
  - Virtual pitch

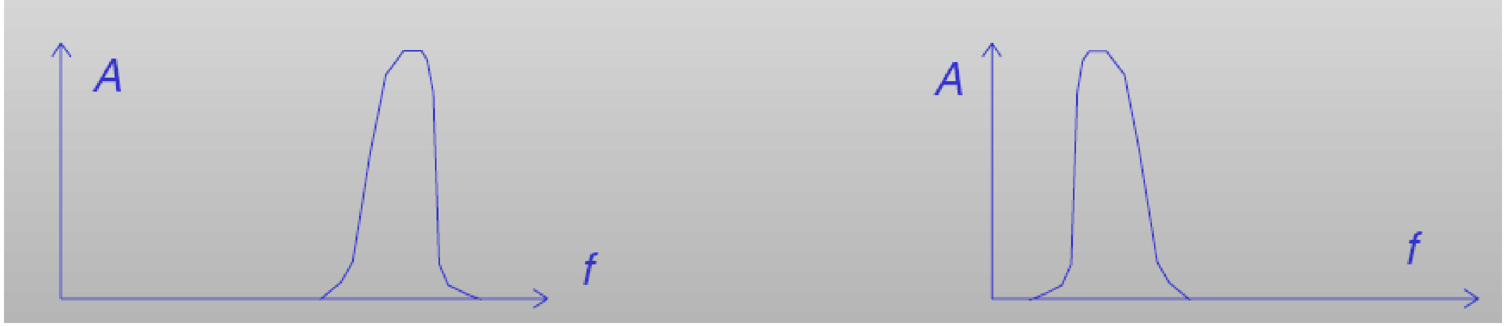
  - **A definition of pitch:**

  attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high (ANSI 1973)

Global height

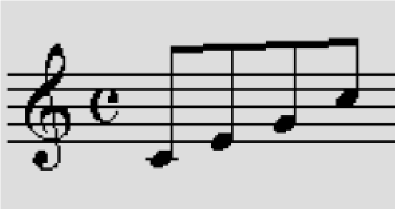
- It is the position of the sound on subjective scale « low-high", independant of any musical sense.

- It is related to the concentration of energy on the frequency axis:



Tonal pitch

- Express the position of sound in a set of intervals
- A serie of intervals = melody



- applicable to pure sounds and periodic sounds

Spectral pitch

- The different height that can be distinguished in a complex sound (analytic listening)
- A set of spectral pitch heard simulatenously may be a chord.
- Perception can be « global » or « analytic »

Virtual pitch

- It is the pitch (perceived height) with a global listening.
- This perceived height may not correspond to a harmonic of the spectrum.
- The ear can hear one or several heights even in non-harmonic sounds .
- … and there is some sounds which have several virtual pitches (for instance bells)

Towards the MEL scale

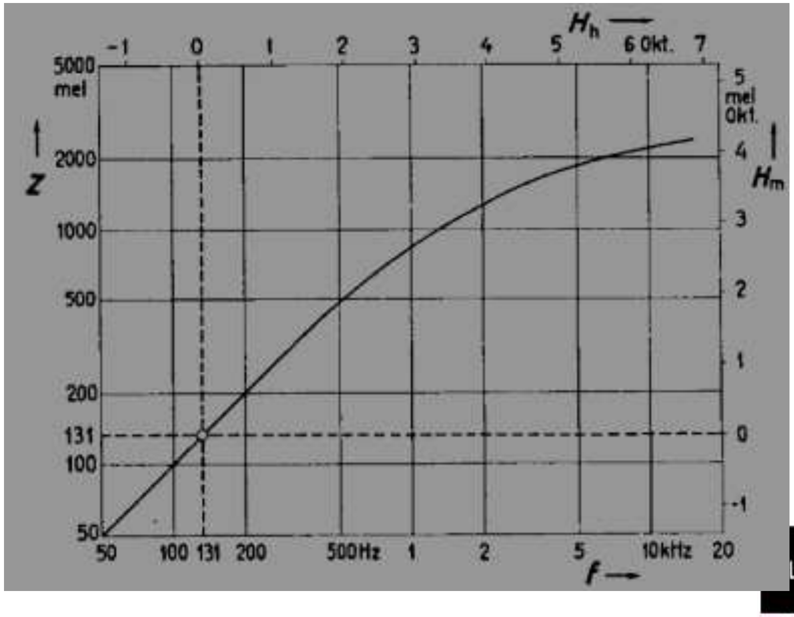
Pitch of pure sounds

- **Experiments:** From a reference sound (sinusoid @ 1kHz), the « tonie » doubles if another sound is perceived twice as high, etc…

- **Results:** Tonie is proportional to frequencies (in Hz) for low frequencies and logarithmic for higher frequencies

⇨ **More precisely, there are two scales**

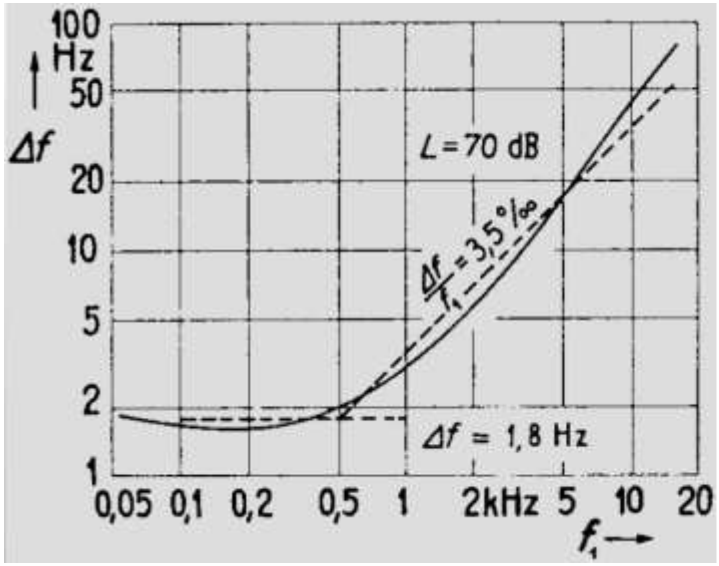
- From 0 to 500 Hz where 1 Mel = 1 Hz (linear)
- above 500 Hz where tonie is a logarithmic function of frequency



Just Noticeable Differences (JND) of pitch

- **Pitch variation is often more important than absolute pitch**
  - Warning function: Doppler effect gives us information on the speed of a sound source
  - Pleasantness: « wowing» of vinyl records

- **Measures of JND give (in laboratory) about 2Hz at 500Hz (500 to 502 Hz)**



Complex sounds

- **Some definitions:**
  - **Complex sounds** are all sounds that are not pure sinusoids
  - **Partials**: frequency component with energy
  - **Harmonics:** are partials in harmonic relation (multiple of a fundamental frequency)
  - **Fundamental frequency:** the first harmonic

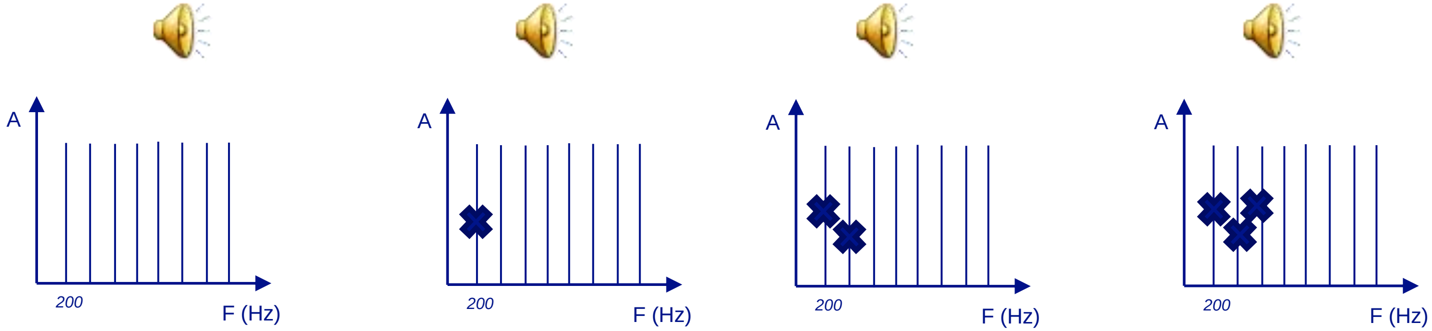
- **In most cases, the ear summarizes the perception of all partials to hear one or several heights**
  - It is in fact the case of harmonic sounds.

- **But**
  - Height perception is not imposed by the lowest partial
  - Height is not imposed by the most energetic part of the spectrum
  - Height is not really independent of « timbre »

The missing fundamental

- **In a complex sound, we usually perceive height even if there are no fundamental frequencies.**

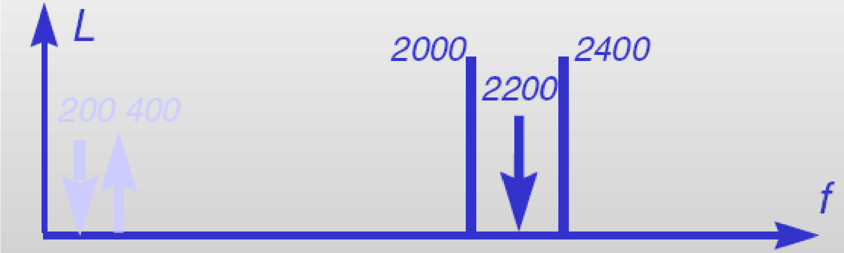
- ***Démo :*** We take out from the low frequency the n harmonics of a sound with a fundamental frequency of 200 Hz.



Octave ambiguity

- **Three partials: 2000 + 2200 + 2400 Hz**
  - We hear a pitch of 200 Hz !

- **If the level of the partial at 2000 Hz decreases progressively :**



  - … at some point the « spectral sight » of this partial becomes too weak and we rather hear a pitch of 400 Hz

- **At some point we hear the two simultaneously: there is an ambiguity of octaves.**

Analytic vs global perception

- **Depending of the sound, the ear perceives differently the set of the sound components:**
  - Either as a set of **distinctive** components
  - Or, **globally,** as 1 or several heights and an associated timbre.

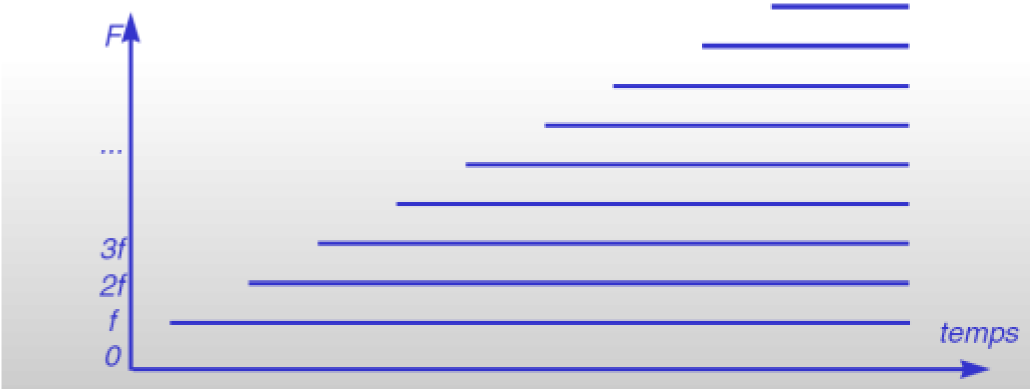
- **Perception depends :**
  - On stimulus,
  - The listener will,
  - On how the stimulus is played (order of arrival/stop of partials)

- **Demo: auditive illusion….**

Circularity in pitch judgement

Analytic vs global perception

Demo **:** analysis / synthesis by progressively adding the partials



Complex sound

harmonic per harmonic

- **Until when the analytic listening is possible ?**

What is the absolute pitch ?

- It is the capacity to recognize and name the height of a musical sound without a sound of reference

- A rare faculty … (less than 1% of the population).

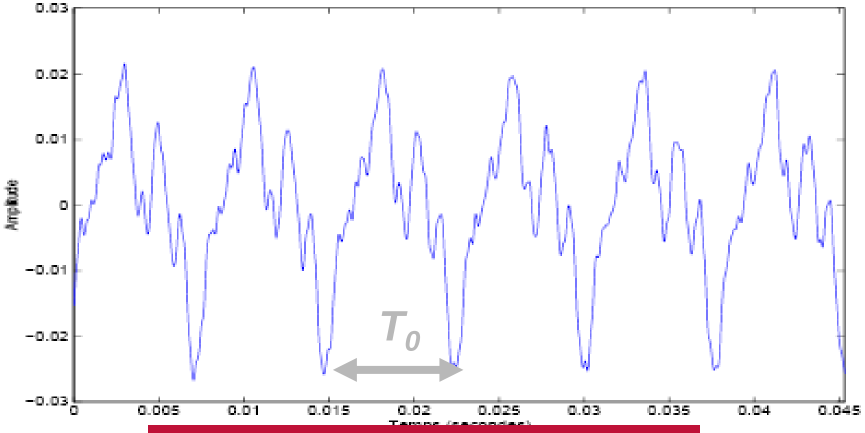
- Can be trainded to some extent… (but results are rarely as good as when the faculty is natural, or acquired when very young…)

Fundamental frequency detection

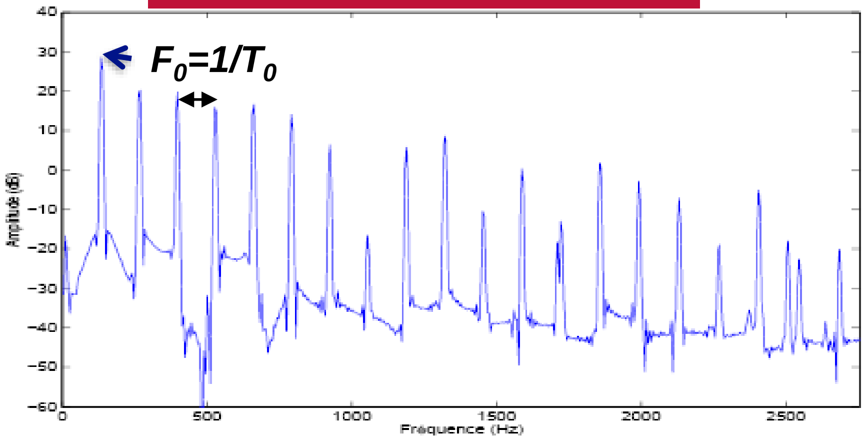
Content

- **Introduction**
  - Quasi-periodic sounds
  - Quasi-periodic model
- **Time-domain methods**
- **Spectral domain methods**
- **Extension to mutipitch (e.g. multiple fundamental frequencies) estimation**

A quasi-periodic sound



A piano sound (C3)



Spectrum of a piano sound

How can we estimate the height (pitch) of a note

or

How to estimate the **fundamental periode** (T_0)
or **frequency** (F_0) ?

Signal Model

- $$x(n) = \sum_{k=1}^{H} 2A_k \cos(2\pi k f_0 n + \phi_k) + w(n)$$
- $f_0 = \dfrac{1}{T_0}$ **normalised fundamental frequency**
- **H is the number of harmonics**
- **Amplitudes {A_k} are real numbers > 0**
- **Phases {ϕ_k} are independant r.v. uniform on [0, 2π [**
- ***w* is a centered white noise of variance σ^2, independent of phases {ϕ_k}**
- **x(n) is a centered second order process with autocovariance**

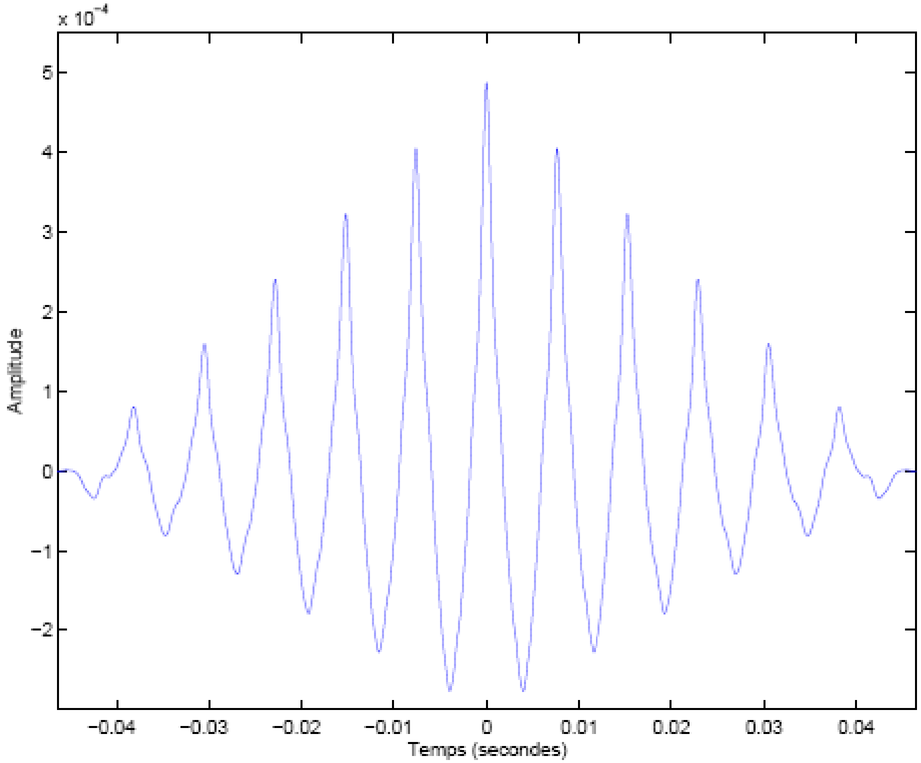
$$r_x(m) = \sum_{k=1}^{H} [2A_k^2 \cos(2\pi k f_0 m)] + \sigma^2 \delta[m]$$

Time domain methods

- **Autocovariance estimation (biased)**

$$\frac{1}{N}\sum_{n=0}^{N-1-m} x[n]\,x[n+m] \text{ si } m \geq 0$$

$$\mathbf{E}(\hat{r}_x[m]) = \frac{N-|m|}{N}\, r_x[m] \qquad\qquad |\hat{r}_x[m]| \leq \hat{r}_x[0]$$

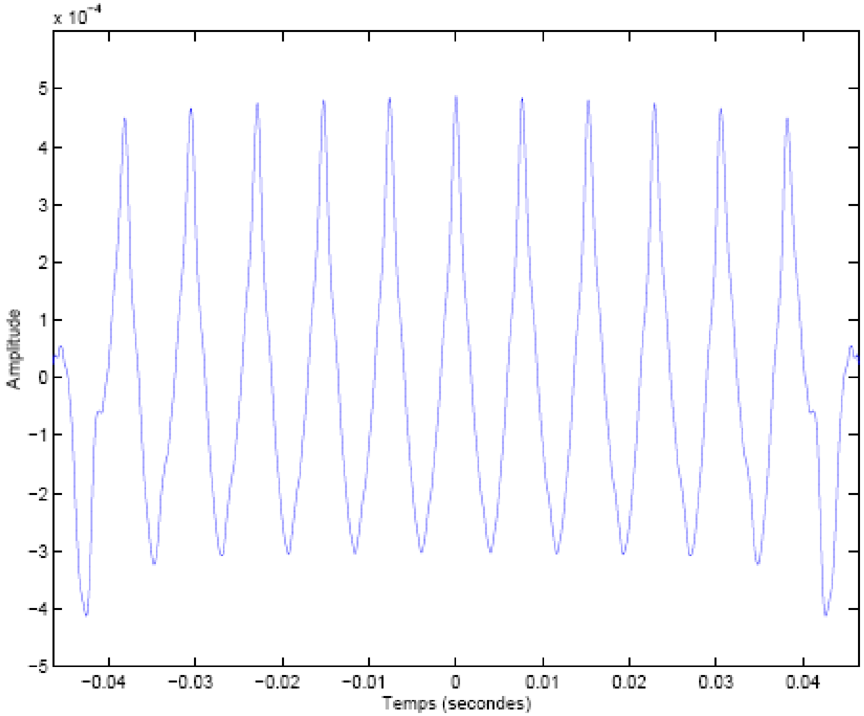


Time domain methods

- **Autocovariance estimation (unbiased)**

$$\tilde{r}_x[m] = \frac{1}{N-m} \sum_{n=0}^{N-1-m} x[n]\, x[n+m] \text{ si } m \geq 0$$

$$\mathbf{E}(\tilde{r}_x[m]) = r_x[m] \qquad \mathrm{Var}(\tilde{r}_x[m]) = \left(\frac{N}{N-m}\right)^2 \mathrm{Var}(\hat{r}_x[m])$$



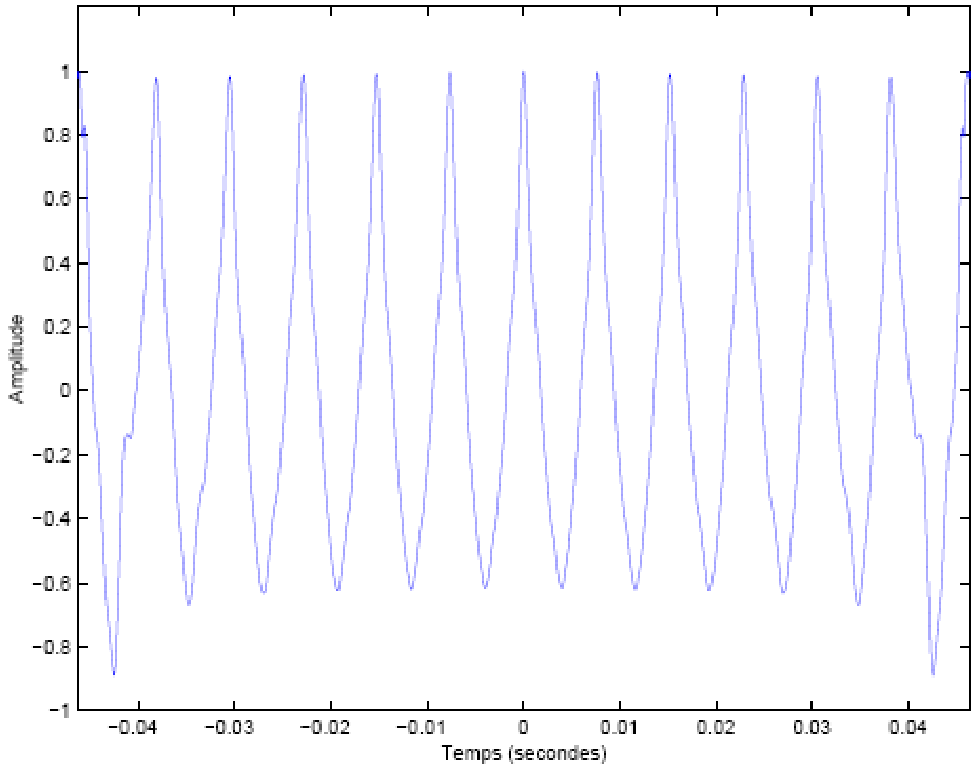
$$|\tilde{r}_x[m]| \nleq \tilde{r}_x[0]$$

Time domain methods

- **Autocorrelation**

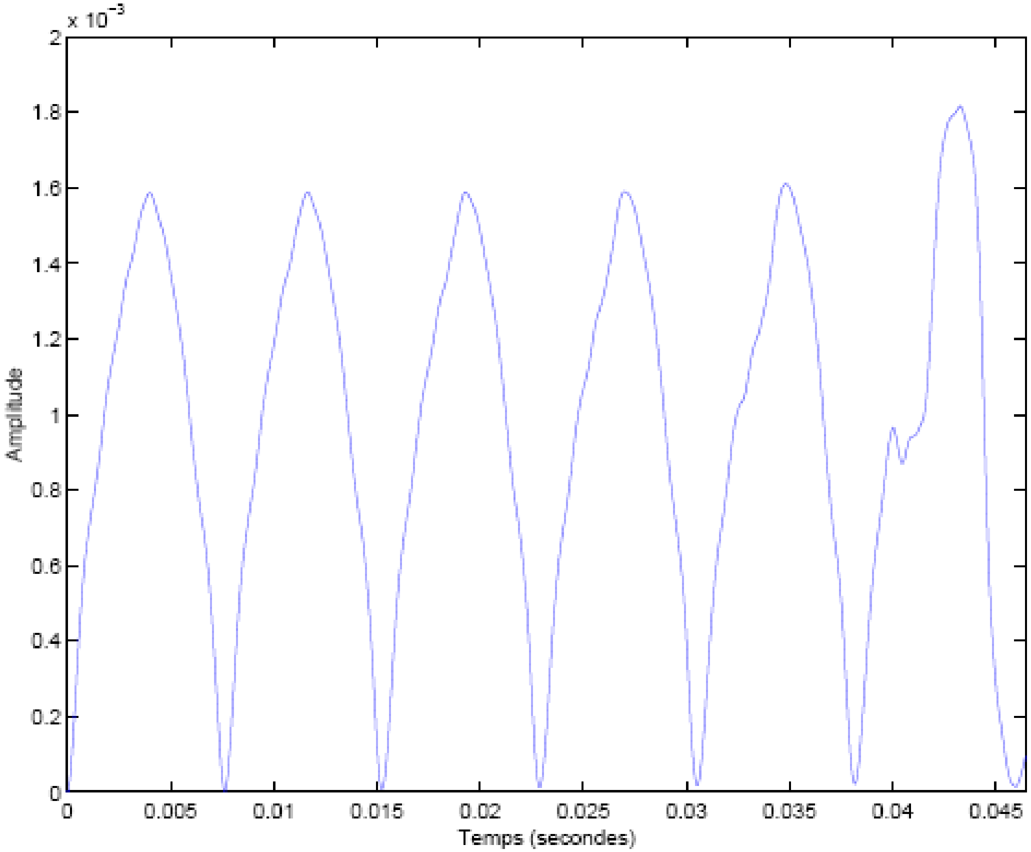
$$\bar{r}_x[m] = \frac{\sum_{n=0}^{N-1-m} x[n]\, x[n+m]}{\sqrt{\sum_{n=0}^{N-1-m} x[n]^2}\sqrt{\sum_{n=0}^{N-1-m} x[n+m]^2}} \text{ si } m \geq 0$$

$$|\bar{r}_x[m]| \leq \bar{r}_x[0] = 1 \qquad |\bar{r}_x[m]| = 1 \text{ ssi les vecteurs sont colinaires}$$



Average square difference function (ASDF)

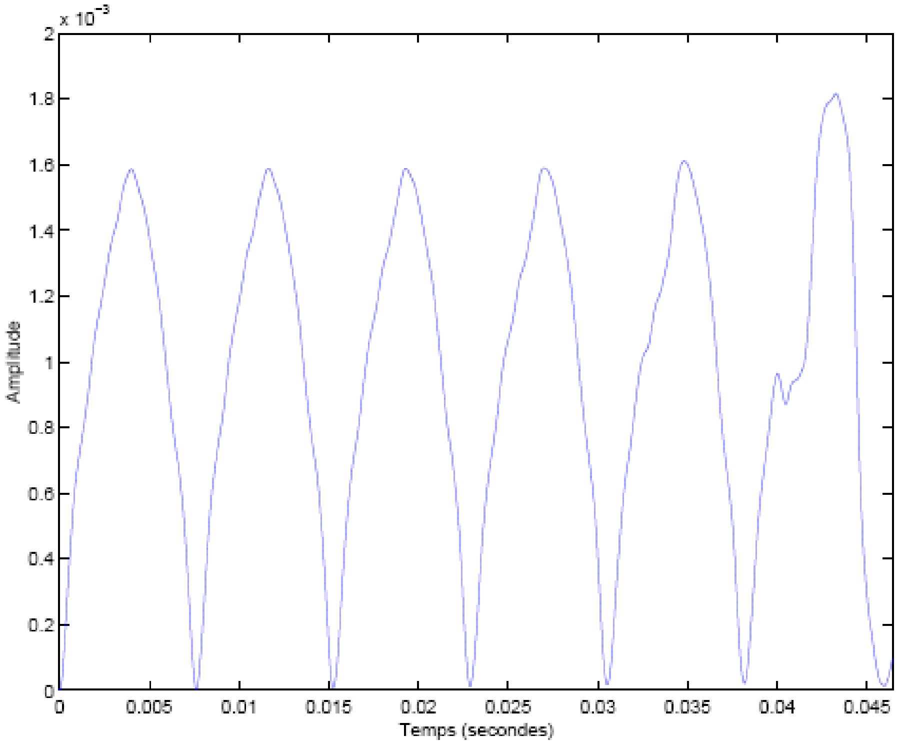
$$\mathrm{ASDF}[m] = \frac{1}{N-m}\sum_{n=0}^{N-1-m}(x[n]-x[n+m])^2$$



Average square difference function (ASDF)

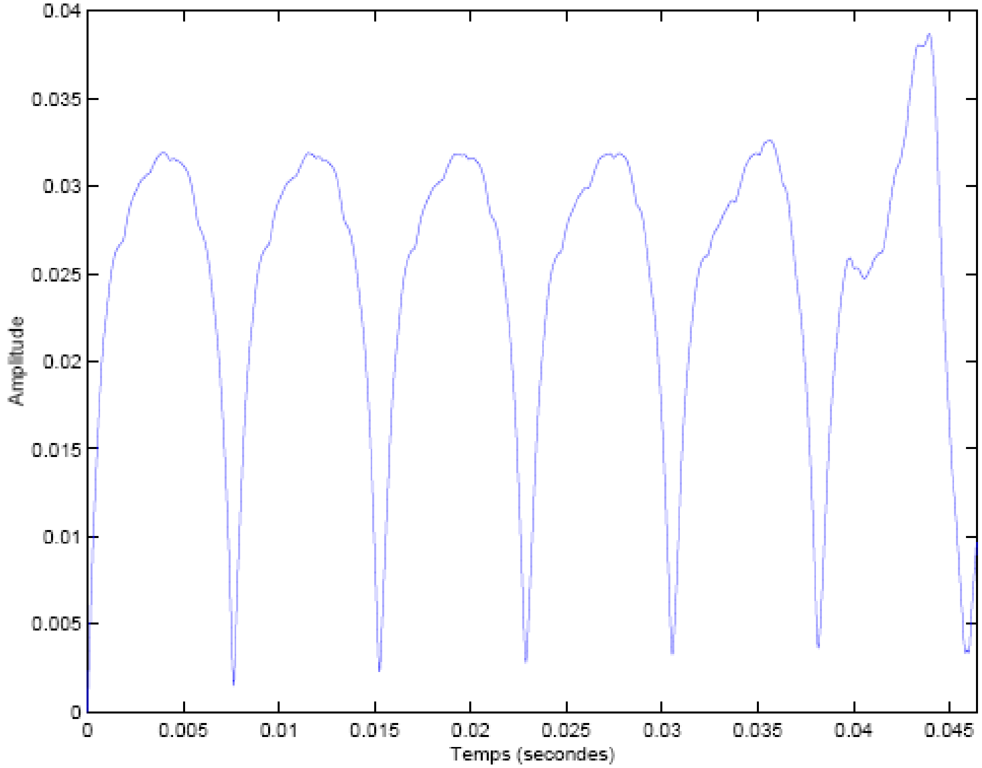
- The period T_0 can be estimated in looking at teh minimum of the square difference between $x(n)$ and $x(n-m)$:

$$\mathbf{E}[\mathrm{ASDF}[m]] = 2(r_x[0] - r_x[m])$$



Average magnitude difference function (AMDF)

$$\mathrm{AMDF}[m] = \frac{1}{N-m} \sum_{n=0}^{N-1-m} |x[n] - x[n+m]|$$

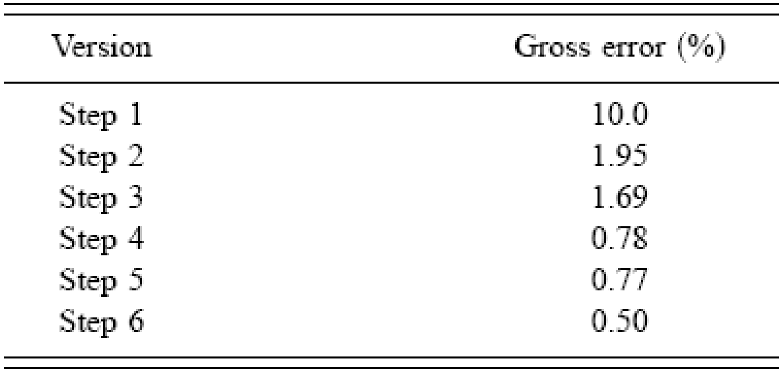


An efficient time-domain algorithm: Yin

(Thanks to V. Emiya for additionnal slides)

- H. Kawahara A. de Cheveigné, *YIN, a fundamental frequency estimator for speech and music*,, JASA, 111(4), 2002

- **Initial method: Autocorrelation method (ACF)**
- **Successive improvements:**
  - Use of ASDF
  - Normalisation
  - Threshold
  - Interpolation
  - Local minimisation in time



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

YIN (2)

- **ASDF used:**

$$\mathrm{d}_n[m] = \sum_{k=0}^{N-1}(x_n[k] - x_n[k+m])^2$$

- **Links with autocorrelation**

$$\mathrm{d}_n[m] = r_n(0) + r_{n+m}(0) - 2r_n(m)$$

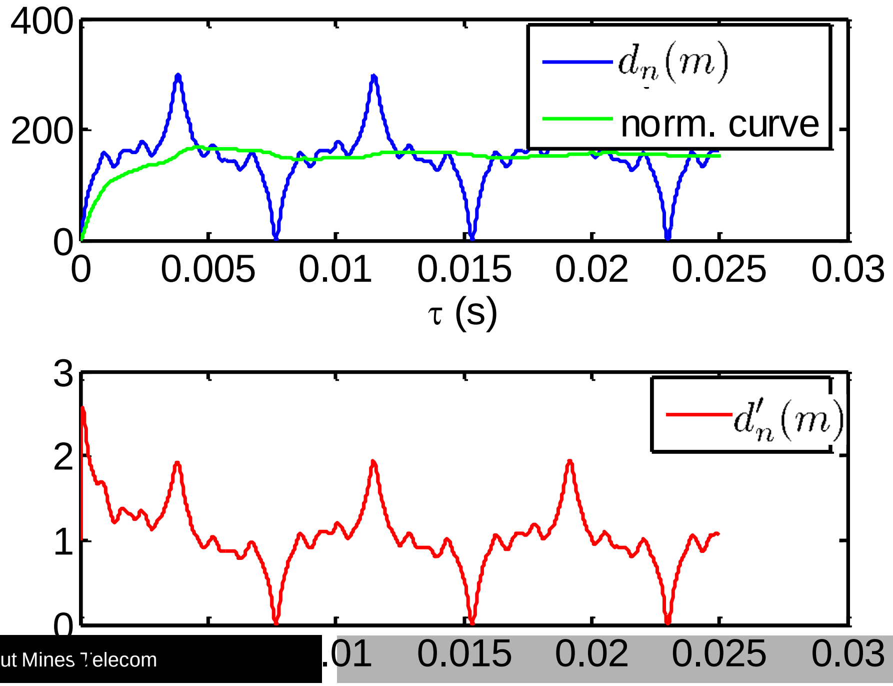
- **Performance increase : ASDF is less sensitive to amplitude variations** (e.g. ACF is sensitive to even harmonics accentuation)

YIN (3)

- **Normalisation by the « cumulative mean »**

$$d'_n(m)(= \begin{cases} 1 & \text{si} \quad m = 0 \\ \dfrac{d_n(m)}{\frac{1}{m}\sum_{k=1}^{m} d_n(k)} & \text{sinon} \end{cases}$$

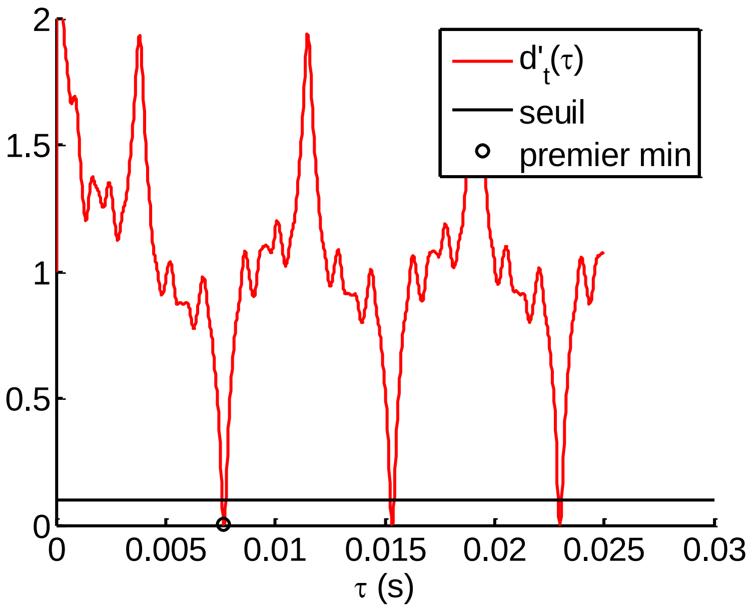
- **Performance increase: suppression of the main lobe at 0**



YIN (4)

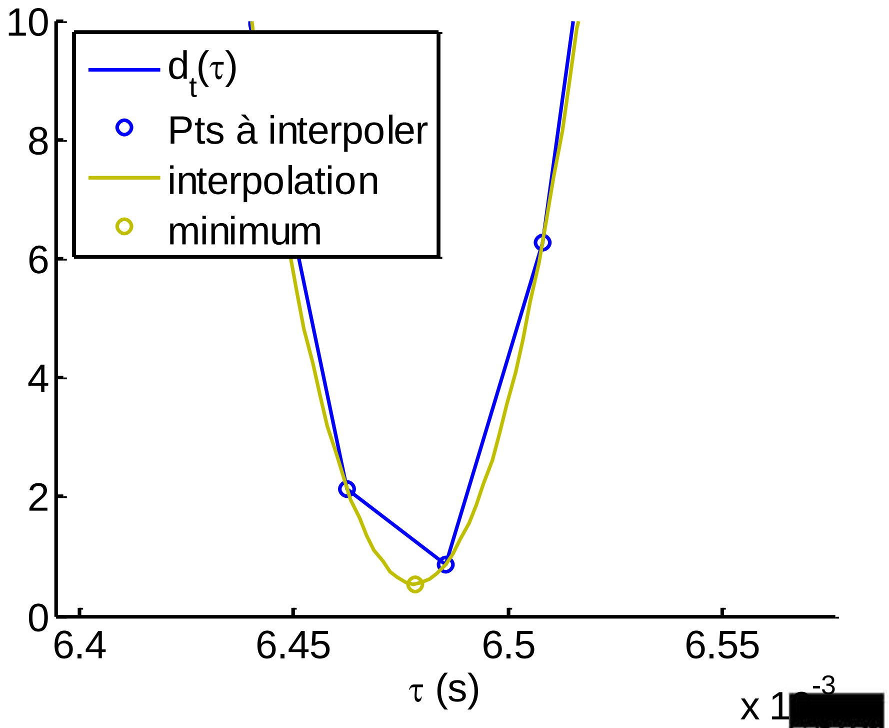
Absolute threshold

- The smallest period below the threshold is chosen
- If no period is below the threshold, the global minimum is chosen



YIN (4)

- **Parabolic interpolation around the minimum**
  - Applied on $d_n(m)$ (i.e before normalisation)
  - Performance increase: precision on F0



YIN (5)

- **Local minimisation in time**

$$T_n = argmin_n(d'_n(m))$$

  - **Minimisation around timeT$_\theta$:** $argmin_\theta(d'_\theta(T_\theta))$ **with**

$$t - T_{max} \quad < \theta \quad < t + T_{max}, \qquad T_{max} = 25ms$$

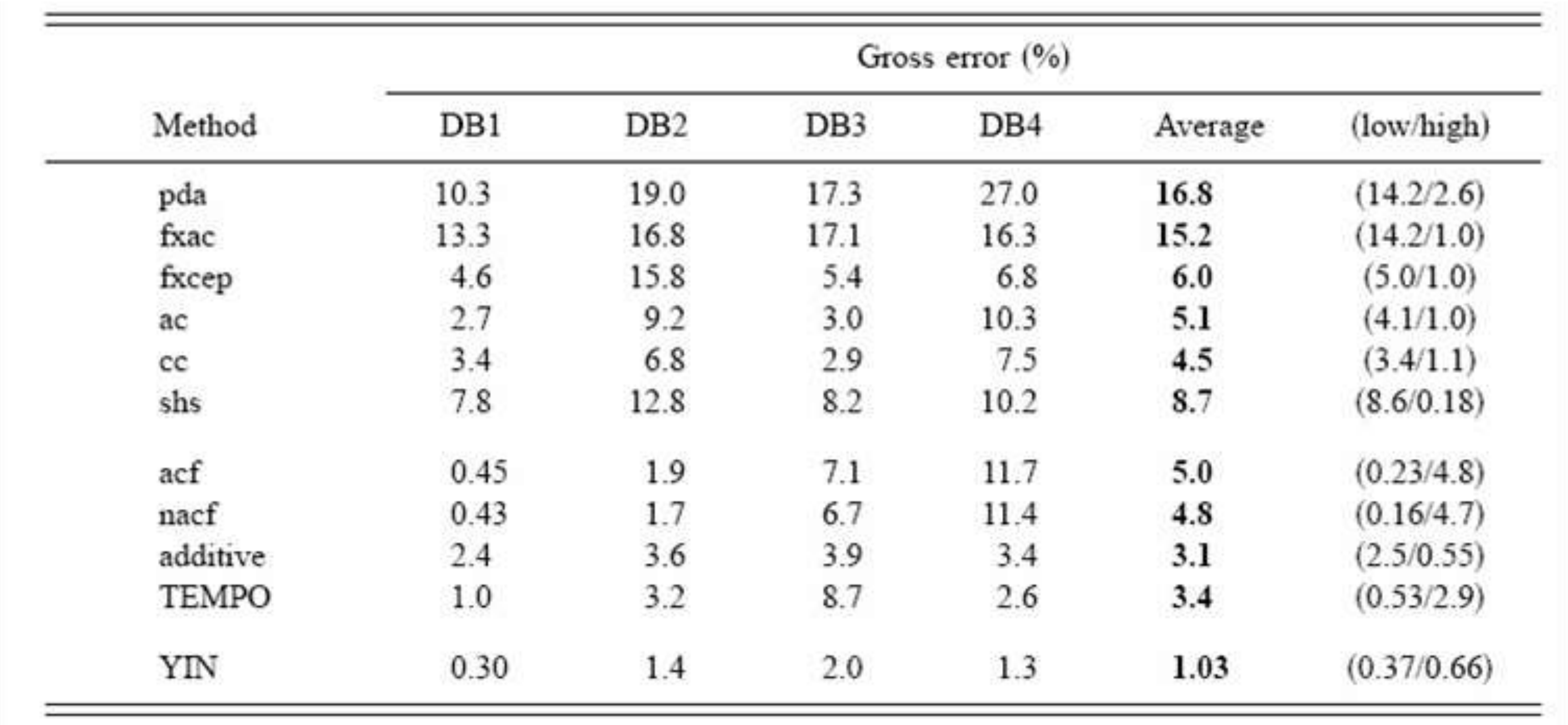
$$0.8T_n \quad < T_\theta \quad < 1.2T_n$$

- **Performance increase in case of fluctuation (it is a kind of smoothing, a bit similar to median filtering)**

Droits d'usage autorisé

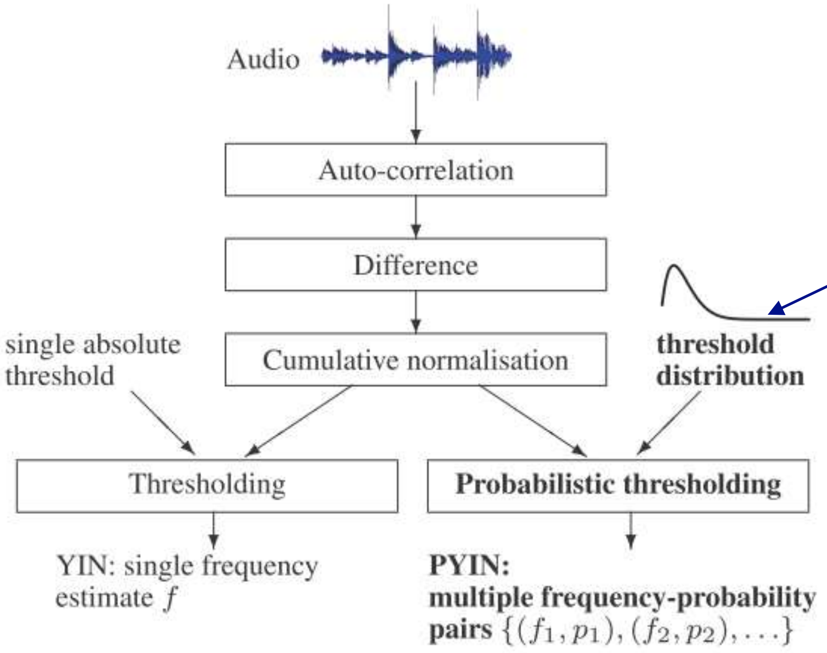
YIN: Evaluation

- **On four speech databases, automatically annotated by YIN (from the laryngograph signal) then manually checked**



	Gross error (%)					
Method	DB1	DB2	DB3	DB4	Average	(low/high)
pda	10.3	19.0	17.3	27.0	**16.8**	(14.2/2.6)
fxac	13.3	16.8	17.1	16.3	**15.2**	(14.2/1.0)
fxcep	4.6	15.8	5.4	6.8	**6.0**	(5.0/1.0)
ac	2.7	9.2	3.0	10.3	**5.1**	(4.1/1.0)
cc	3.4	6.8	2.9	7.5	**4.5**	(3.4/1.1)
shs	7.8	12.8	8.2	10.2	**8.7**	(8.6/0.18)
acf	0.45	1.9	7.1	11.7	**5.0**	(0.23/4.8)
nacf	0.43	1.7	6.7	11.4	**4.8**	(0.16/4.7)
additive	2.4	3.6	3.9	3.4	**3.1**	(2.5/0.55)
TEMPO	1.0	3.2	8.7	2.6	**3.4**	(0.53/2.9)
YIN	0.30	1.4	2.0	1.3	**1.03**	(0.37/0.66)

PYIN: An extension of YIN



Audio

Auto-correlation

Difference

Cumulative normalisation

single absolute threshold

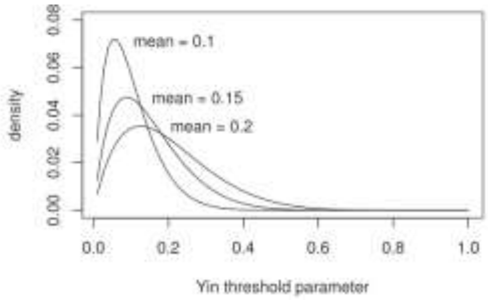
threshold distribution

Thresholding

Probabilistic thresholding

YIN: single frequency estimate f

PYIN: multiple frequency-probability pairs $\{(f_1, p_1), (f_2, p_2), \ldots\}$



$$P(\tau = \tau_0 | S, x_t) = \sum_{i=1}^{N} a(s_i, \tau)\, P(s_i)\, [Y(x_t, s_i) = \tau]$$

$$a(s_i, \tau) = \begin{cases} 1, & \text{if } d'(\tau) < s_i \\ p_a, & \text{otherwise.} \end{cases}$$

$$p_a = 0.01$$

Threshold distribution

Pitch estimate by YIN

$$\begin{aligned} [.] &= 1 \quad \text{if true} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," *2014 IEEE ICASSP*, Florence, Italy, 2014, pp. 659-663, doi: 10.1109/ICASSP.2014.6853678.

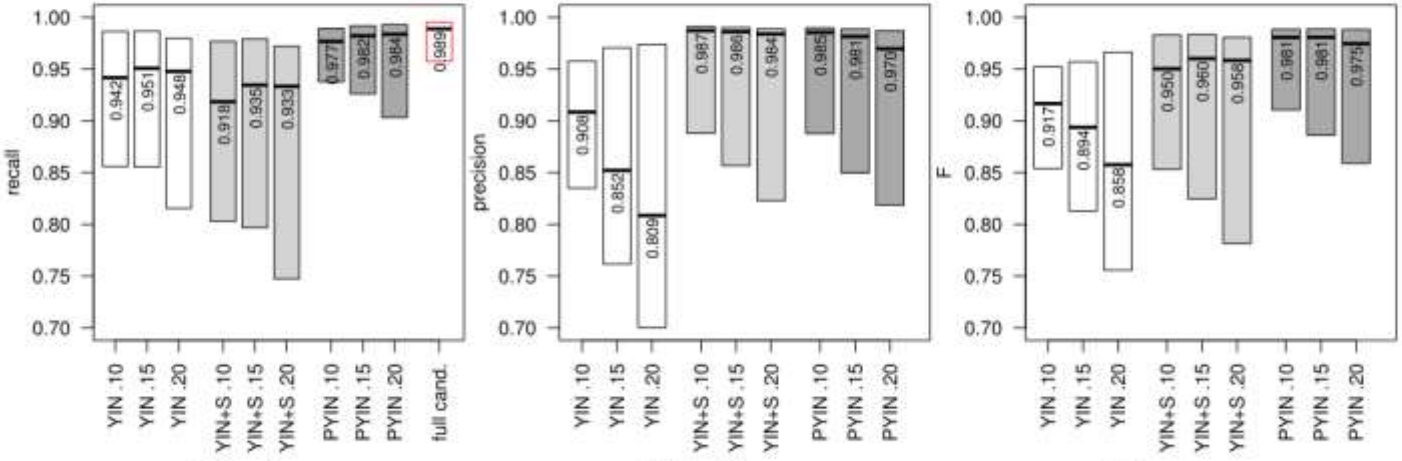
PYIN: An extension of YIN (2/2)

- **Final results obtained by smoothing (using HMM) the set of pairs** $\{f_1, p1), (f_2, p2), ...\}$

- **Some results**
  - Recall = proportion of actually voiced frames which the extractor recognises as voiced and tracks with the correct frequency
  - Precision = proportion of correct pitch estimates in frames marked by the

extractor as voiced

  - F-measure : $F = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$



(a) Recall (b) Precision (c) F measure

Fundamental frequency estimation using a signal model: Maximum likelihood approach

- **Signal model:** $x(n) = a(n) + w(n)$
  - *a* is a deterministic model of period T_0
  - *w* is a Gaussian white noise with variance σ^2

- **Observation likelihood**

$$p(x|T_0, a, \sigma^2)) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x(n)-a(n))^2}$$

- **Log-likelihood**

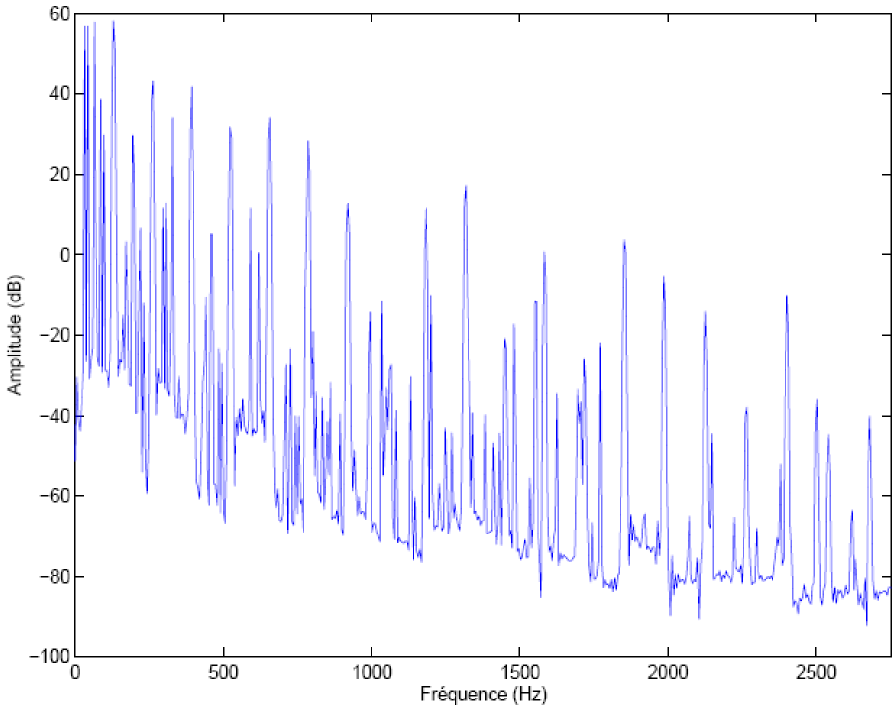
$$L(T_0, a, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x(n) - a(n))^2$$

- **Method :** maximise iteratively L with respect to a, then σ^2 and then T_0

Maximum likelihood approach

- It can be shown that maximisation of L with respect to $F_0 = \frac{m}{N}$ is equivalent to maximise the spectral sum

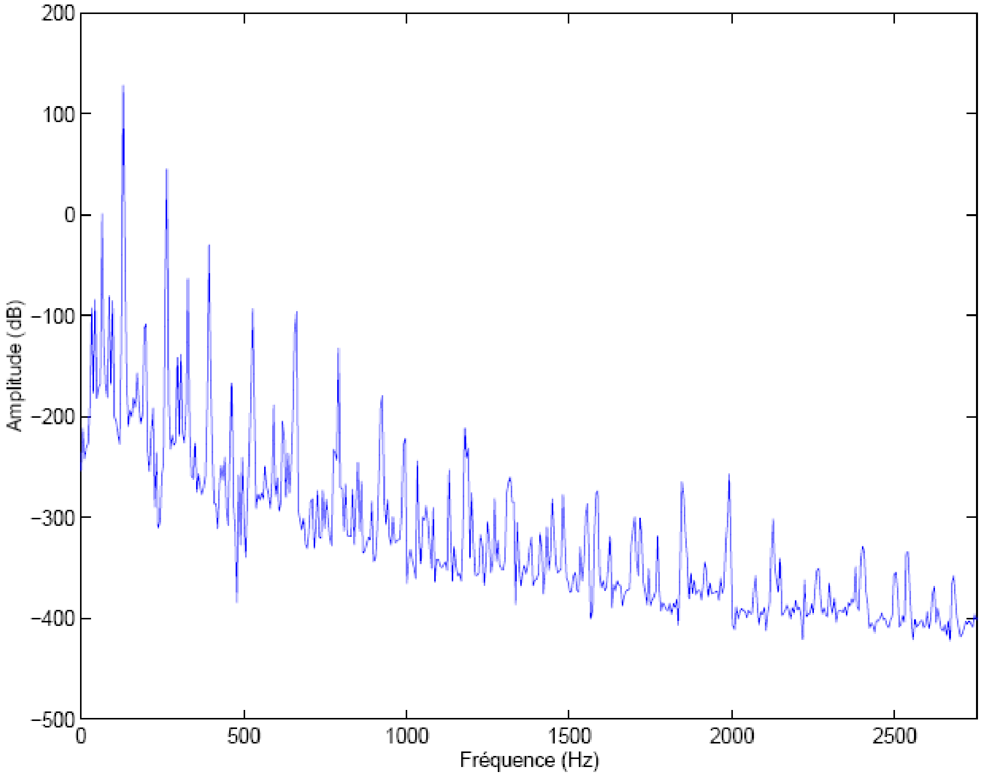
$$S(e^{j2\pi \frac{m}{N}}) = \sum_{k=1}^{H} \hat{R}_x(e^{j2\pi k \frac{m}{N}})$$



Spectral product

- **By analogy to spectral sum (often more robust)**

$$P(e^{j2\pi\frac{m}{N}}) = \prod_{k=1}^{H} \hat{R}_x(e^{j2\pi k\frac{m}{N}})$$



Multiple fundamental frequencies detection

Droits d'usage autorisé

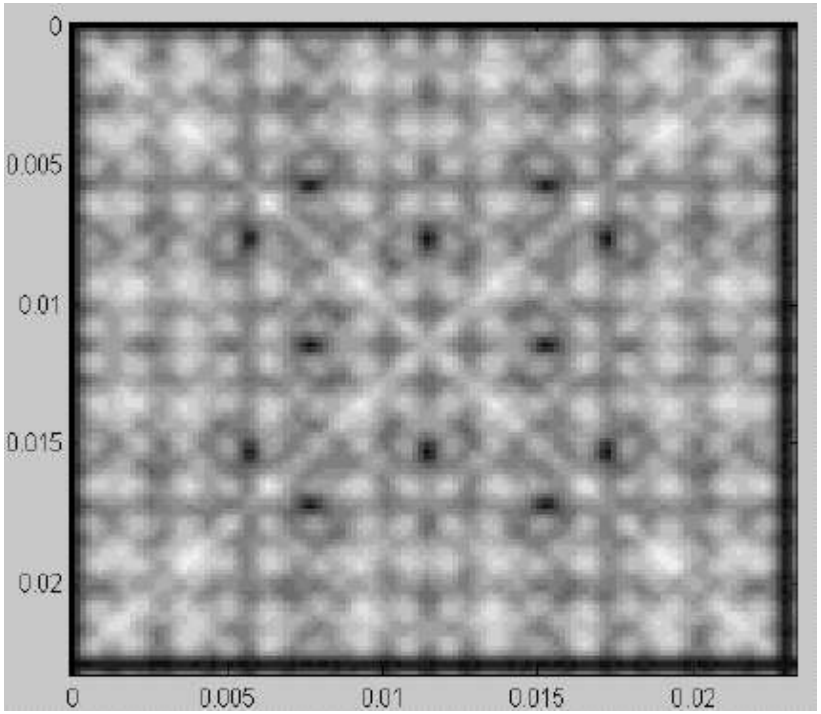
Multiple fundamental frequencies detection

- Objective: to estimate all musical notes of a polyphonic recording
- Problem: notes can be played in harmony (often the case in music …!!)
- Sometimes: necessity to take into account the non-harmonicity of played notes

Multiple fundamental frequencies detection

- **DMDF (*Double Magnitude Difference Function*)**

$$D(k_1,k_2)=\frac{1}{N-k_1-k_2}\sum_{n=0}^{N-k_1-k2-1}|d(n)-d(n+k_1)-d(n+k_2)+d(n+k_1+k2)|$$

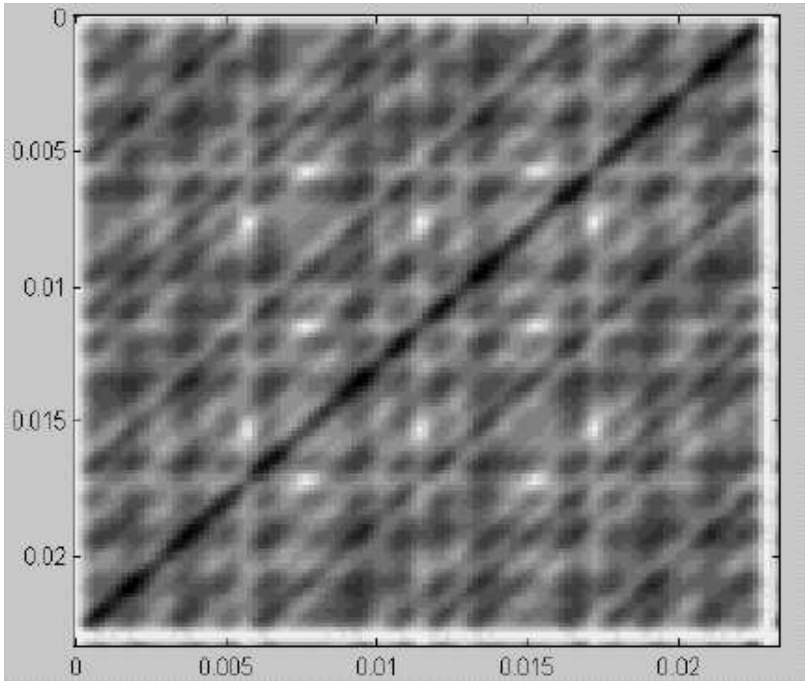


- ✓ **piano sound**
- ✓ addition of two notes
  T1=0.0076s
  T2=0.0057s

Multiple fundamental frequencies detection

- **Bi-dimensional correlation**

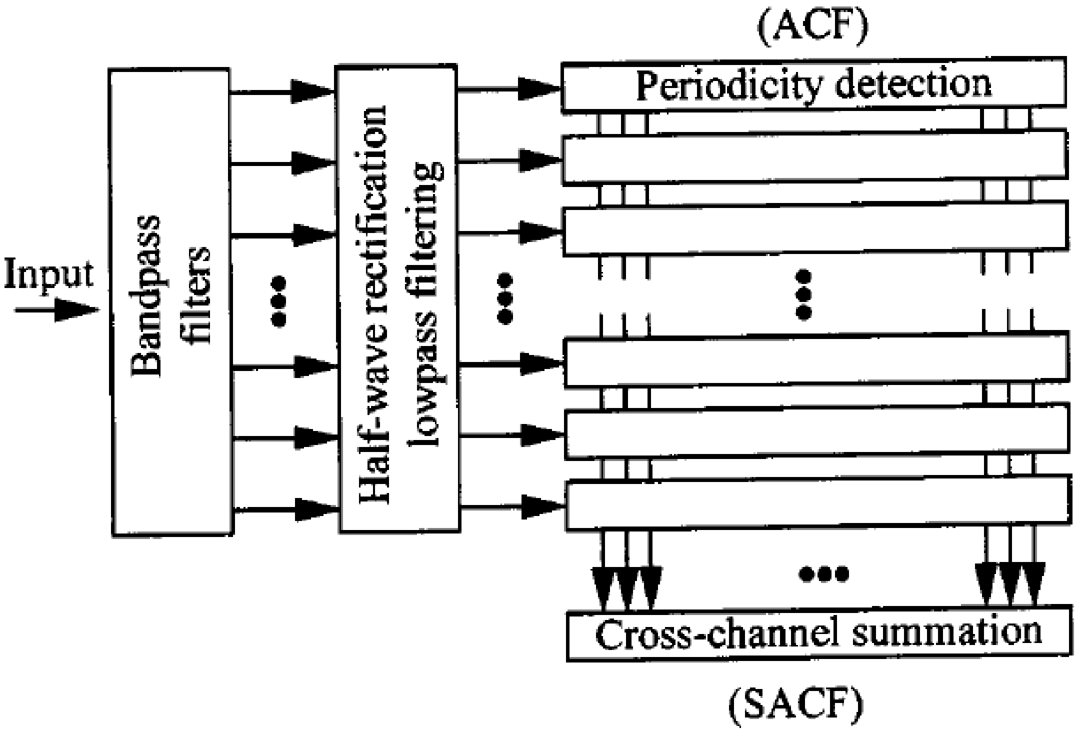
$$\overline{r}\left(k_1,k_2\right)=\frac{\sum_{n=0}^{N-k_1-k_2-1} d\left[n\right]\left(d\left[n+k_1\right]+d\left[n+k_2\right]-d\left[n+k_1+k_2\right]\right)}{\left(\sum_{n=0}^{N-k_1-k_2-1} d\left[n\right]^2\right)^{1/2}\left(\sum_{n=0}^{N-k_1-k_2-1}\left(d\left[n+k_1\right]+d\left[n+k_2\right]-d\left[n+k_1+k_2\right]\right)^2\right)^{1/2}}$$



Measures the « similarity » between

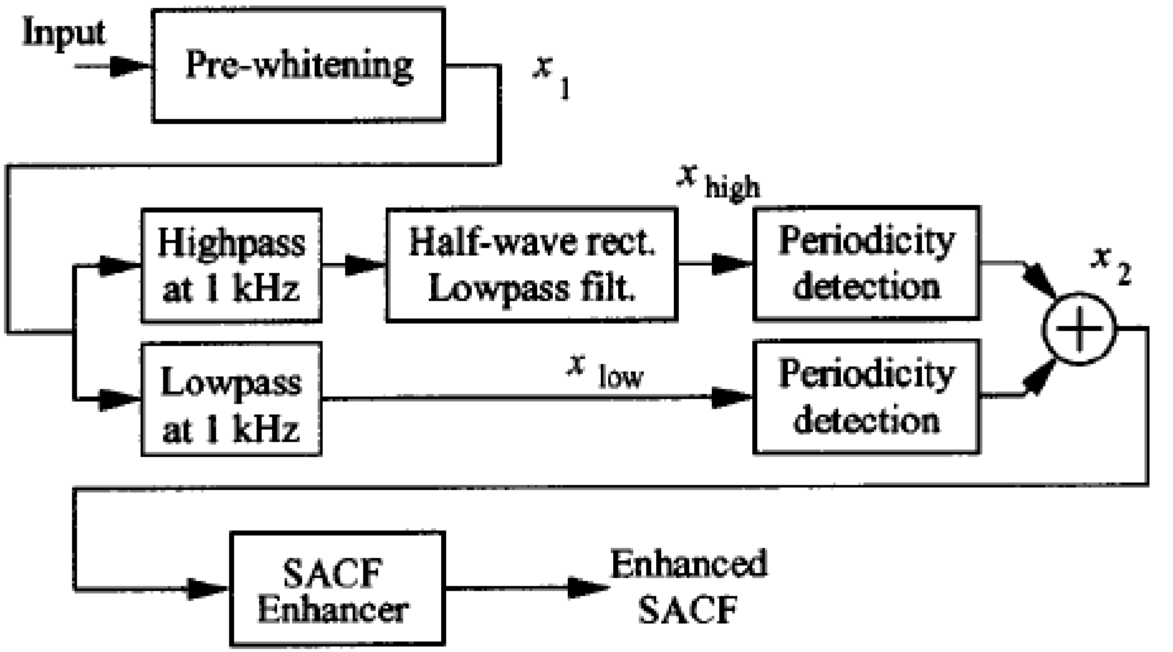
- d(n) et
- d(n+k1) + d(n+k2)-d(n+k1+k2)

A filter bank approach



- R. Meddis and M. Hewitt, “Virtual pitch and phase sensitivity of a computer model of the auditory periphery—I: Pitch identification,” *J. Acoust. Soc. Am.*, vol. 89, pp. 2866–2882, June 1991.

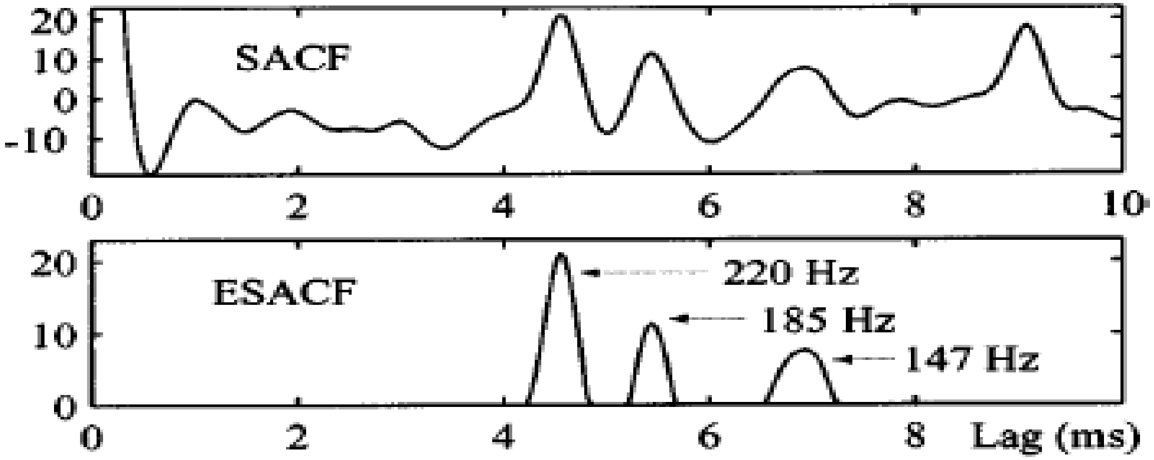
A simpler approach (inspired by the previous method)



- T. Tolonen and M. Karjalainen, “A computationally efficient multipitch analysis model,” *IEEE Trans. On Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

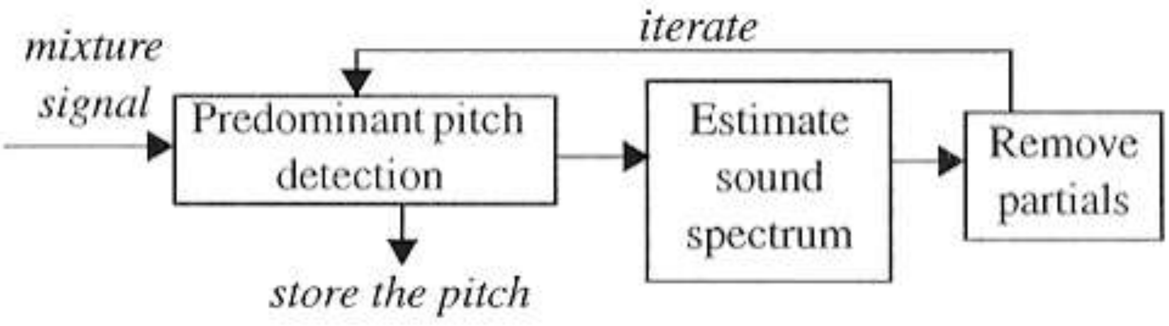
Enhanced Summary ACF

- **Several steps:**
  - **Half wave rectification**
    - We only keep positive values
  - **Slowed down twice (or more) and deduced from rectified SACF**
    - allows to suppress double pics



An iterative approach

- **Estimate each note one after the other …**
  - First, detect the most prominent note …
  - Subtract this note from the polyphony
  - Then, detect the next most prominent note
  - Subtract this note from the polyphony
  - Etc… until all notes are found

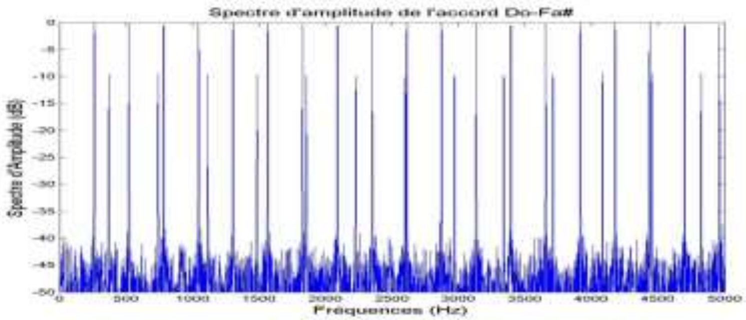


Anssi P. Klapuri, *Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness*, IEEE Trans. On Speech and Sig. Proc., 11(6), 2003

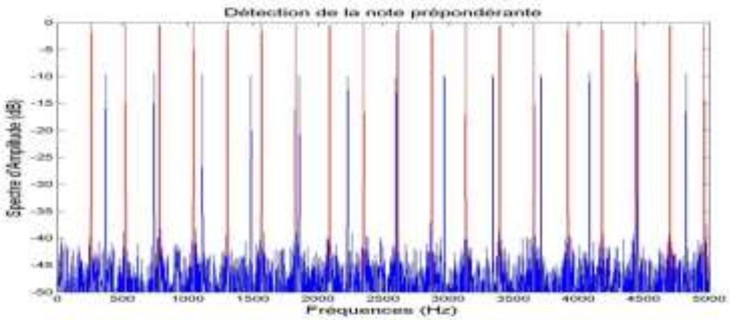
Anssi P. Klapuri “Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model”, IEEE Trans. On ASLP, Feb. 2008

Iterative multipitch estimation

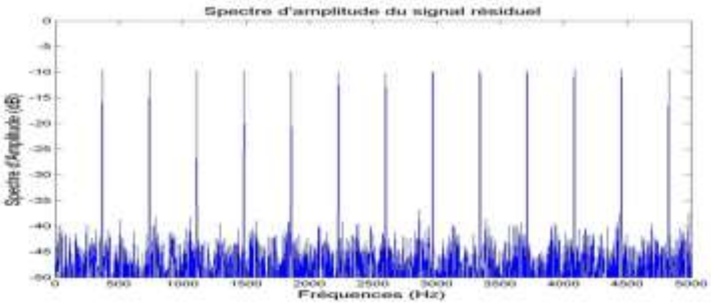
Chord of two synthetic notes C – F#



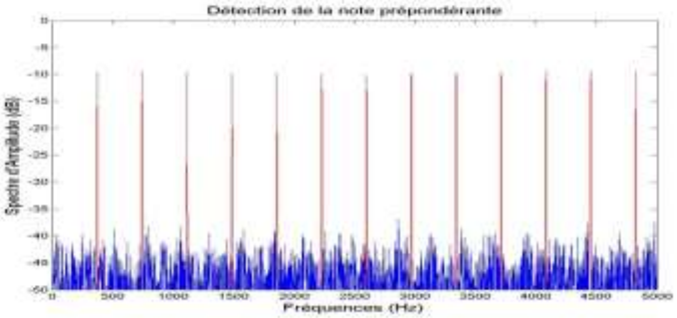
Detect the most prominent note (in red)



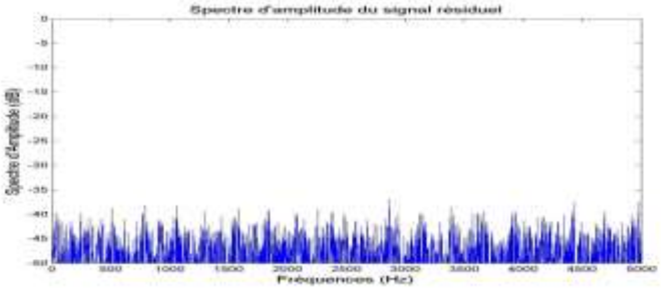
Subtract the detected note



Detect the next most prominent note



There is no more notes….chord C – F# is recognized

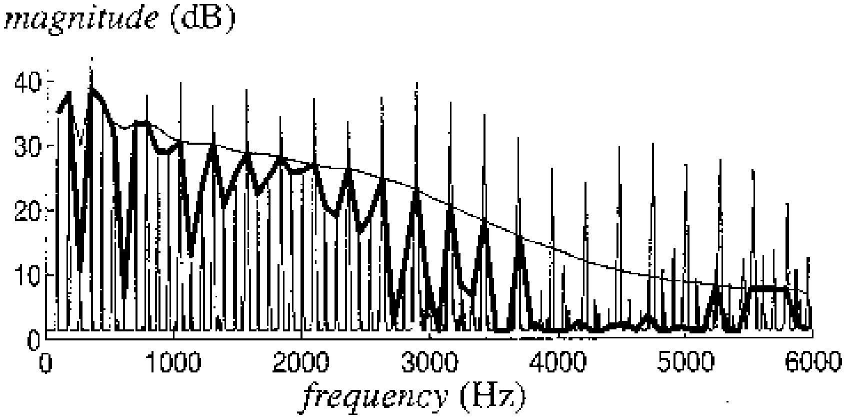


Iterative multipitch estimation

Spectral smoothing: towards subtracting only the current note

- $a_h = min(a_h, m_h)$

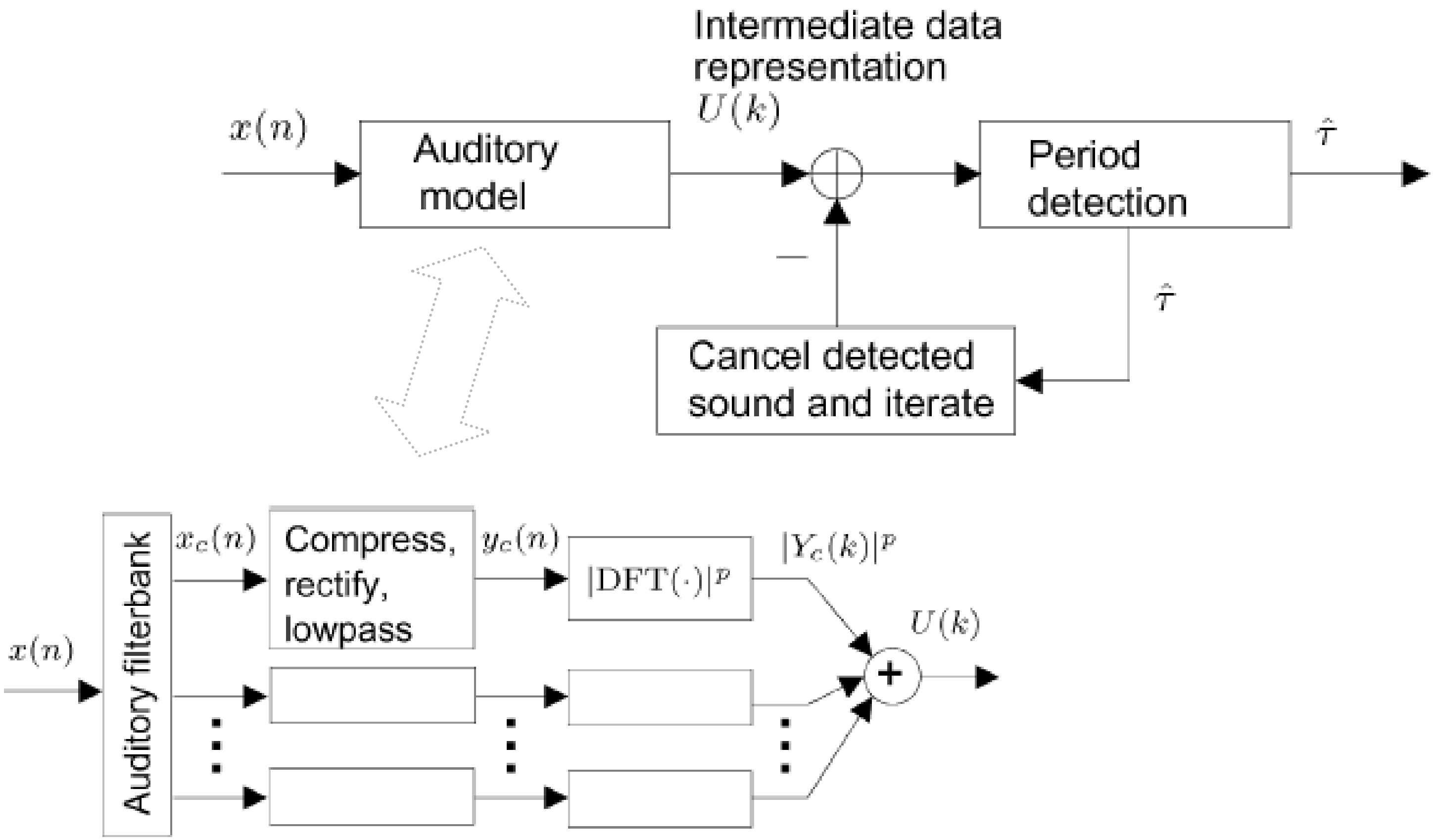
where m_h is the mean on a spectral window (*one octave wide*) around the current harmonic.



Anssi P. Klapuri, *Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness*, IEEE Trans. On Speech and Sig. Proc., 11(6), 2003

Anssi P. Klapuri "Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model", IEEE Trans. On ASLP, Feb. 2008

Improvement using a perceptual model

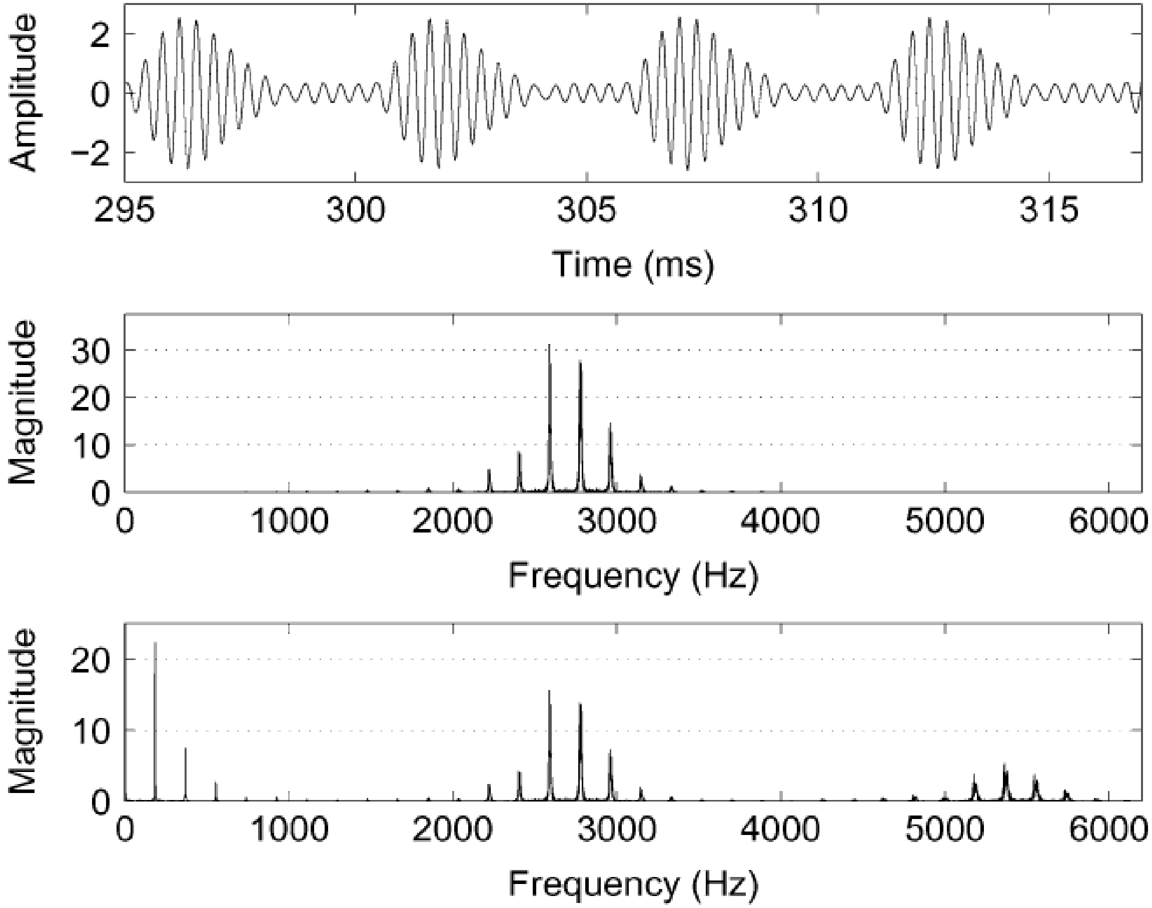


Intermediate data representation $U(k)$
$x(n)$
Auditory model
Period detection
$\hat{\tau}$
$-$
$\hat{\tau}$
Cancel detected sound and iterate
Auditory filterbank
$x(n)$
$x_c(n)$
Compress, rectify, lowpass
$y_c(n)$
$|\mathrm{DFT}(\cdot)|^p$
$|Y_c(k)|^p$
$U(k)$
+

- Anssi P. Klapuri “Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model”, IEEE Trans. On ASLP, Feb. 2008

Effect of compression and rectification

- **Result on a band centered at 2.7 kHz**



Multiple frequency estimation

- **Many other approaches**
  - **Bayesian methods**
  - **Factorisation methods (NMF for example)**
  - **Neural networks, Deep neural networks**

Non-Negative Factorization methods or NMF

- Use of non-supervised decomposition methods (for example Non-Negative Factorization methods or NMF)

- **Principle of NMF :**

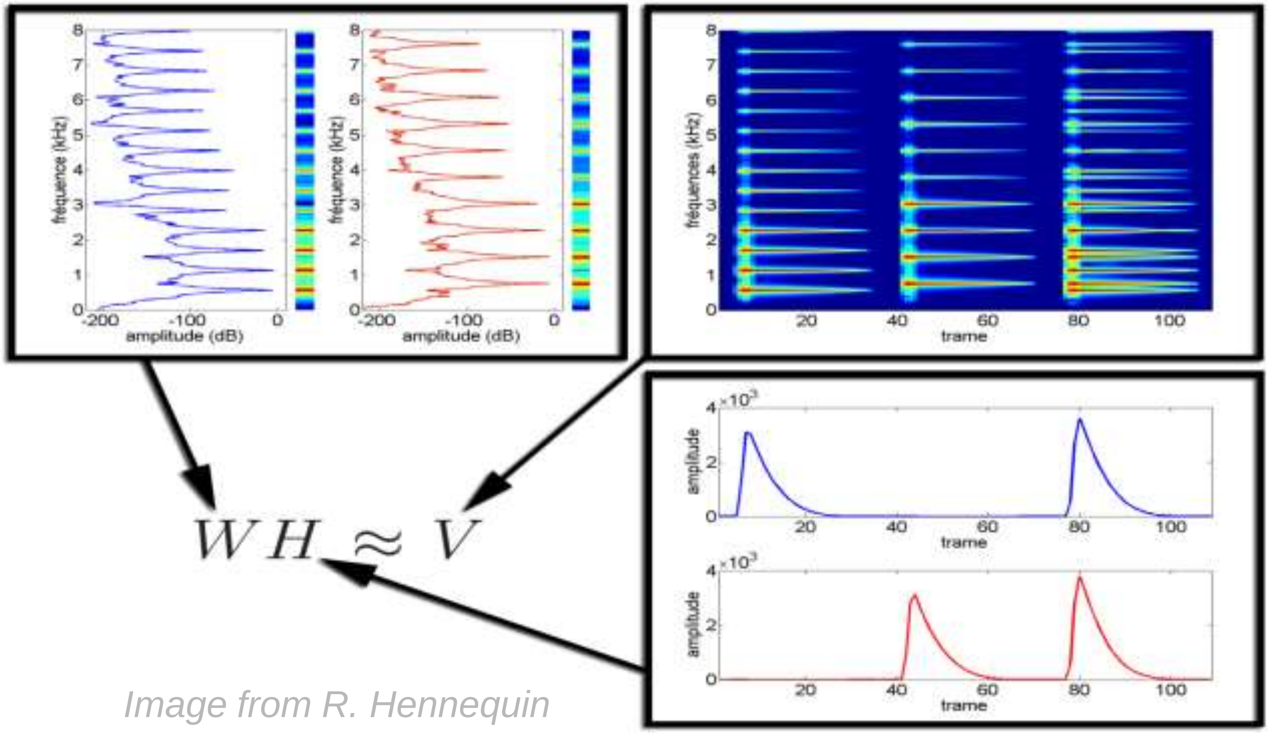


Image from R. Hennequin

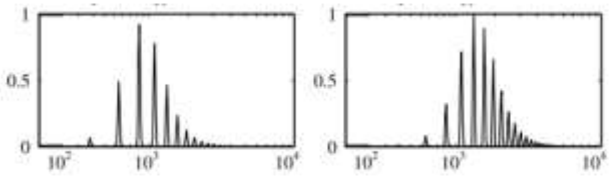
Non-Negative Factorization methods or NMF

Use in multipitch estimation:

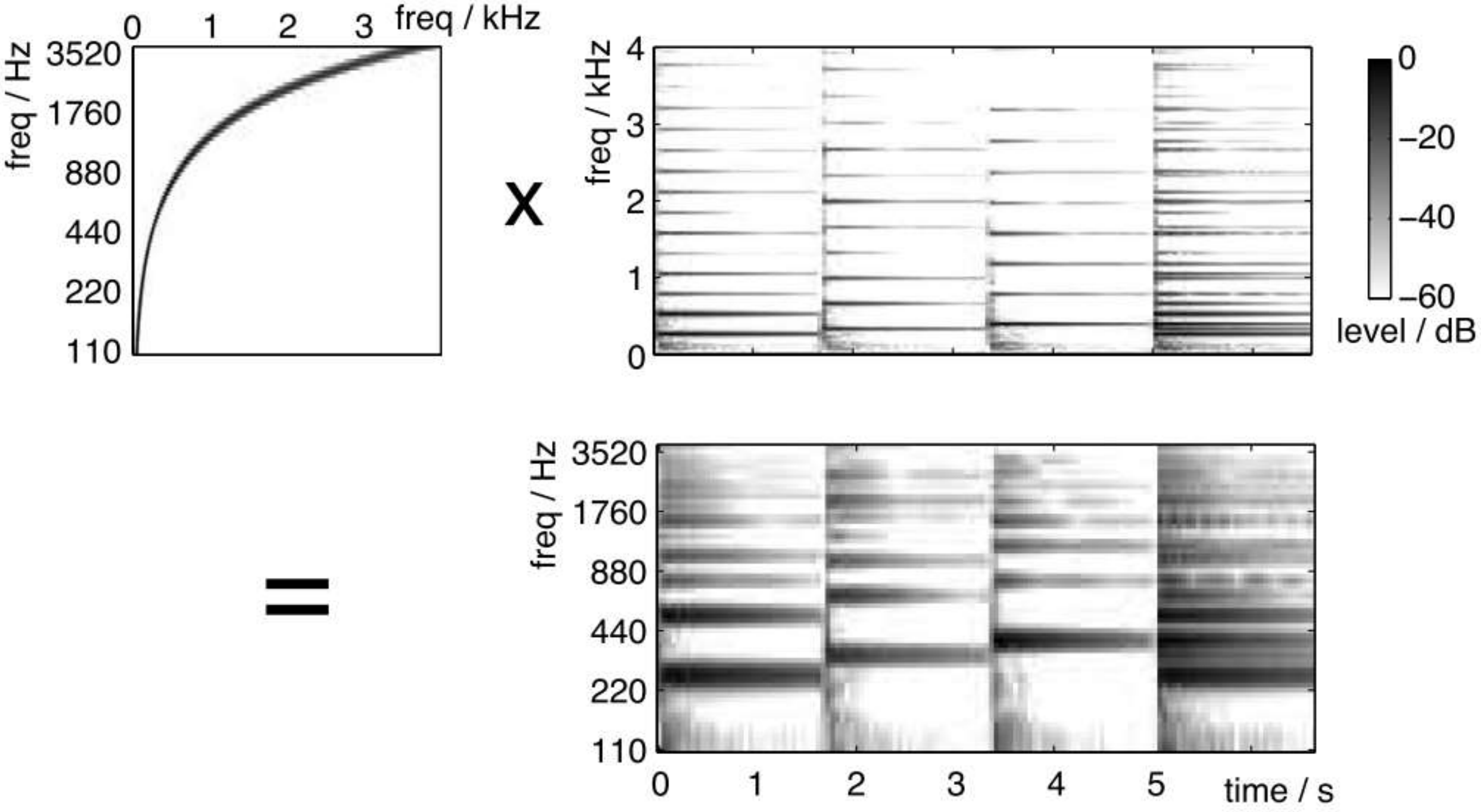
- Important to introduce *a priori* (probabilist approach) or constraints (déterminist approach)

- Constraint examples (after Vincent & al, 2010):

  - NMF classic: $$Y_{ft} = \sum_{i=1}^{I} A_{it} S_{if}$$
  - NMF with pitch dependant templates: $$Y_{ft} = \sum_{p=p_{\mathrm{low}}}^{p_{\mathrm{high}}} \sum_{j=1}^{J_p} A_{pjt} S_{pjf}$$
  - … and template constraints $$S_{pjf} = \sum_{k=1}^{K_p} E_{pjk} N_{pkf}$$
  - Ex. With “local” envelopes



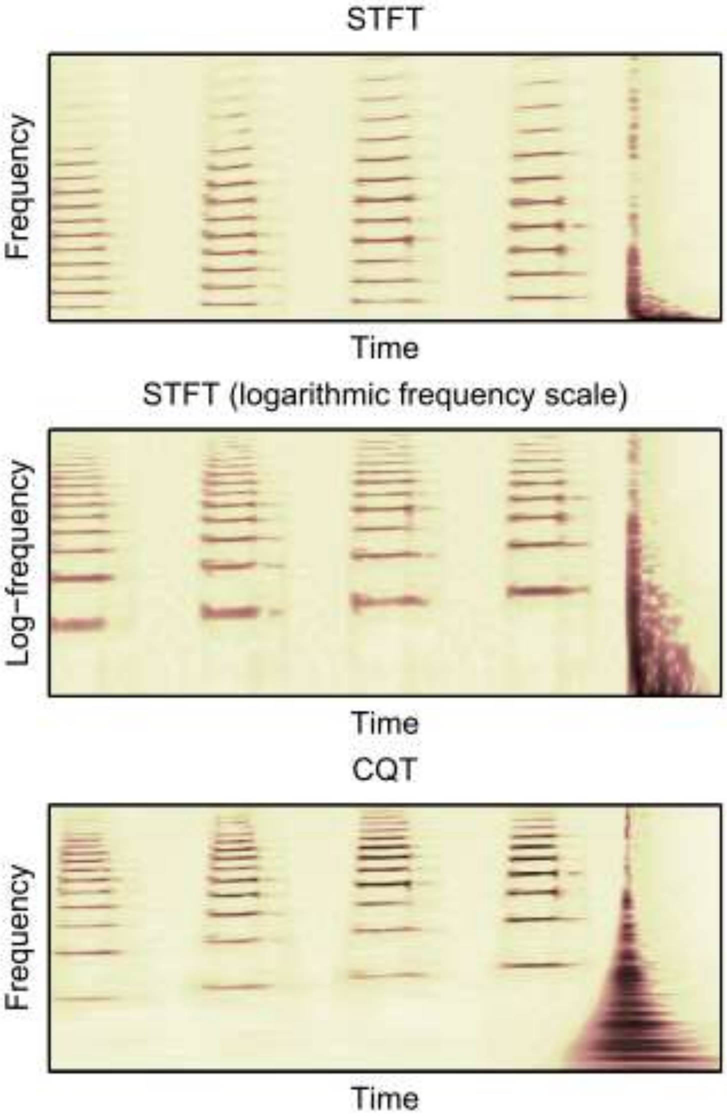
Use of a constant Q transform



D'après M. Mueller & al. « Signal Processing for Music Analysis, IEEE Trans. On Selected topics of Signal Processing, oct. 2011

Utilisation en estimation multipitch

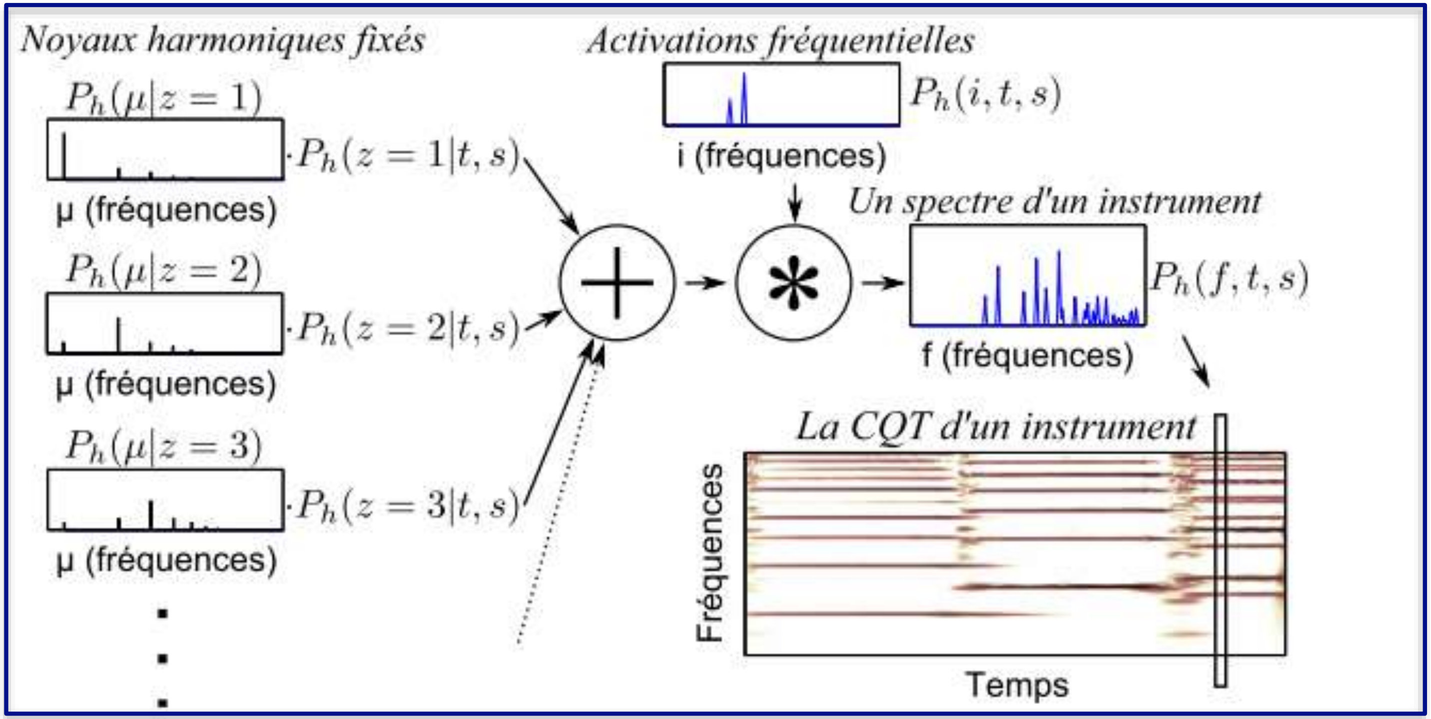
- **On a constant Q transform**
  - **A difference in pitch corresponds to a translation in frequency**
  - **Towards "Shift invariant PLCA (v. smaragdis2008 et Fuentes & al. 2011)**



A PLCA model example

- **The HALCA model (Fuentes & al.)**

$$P(f,t) = P(c=h)\boxed{P_h(f,t)} + P(c=b)P_b(f,t)$$

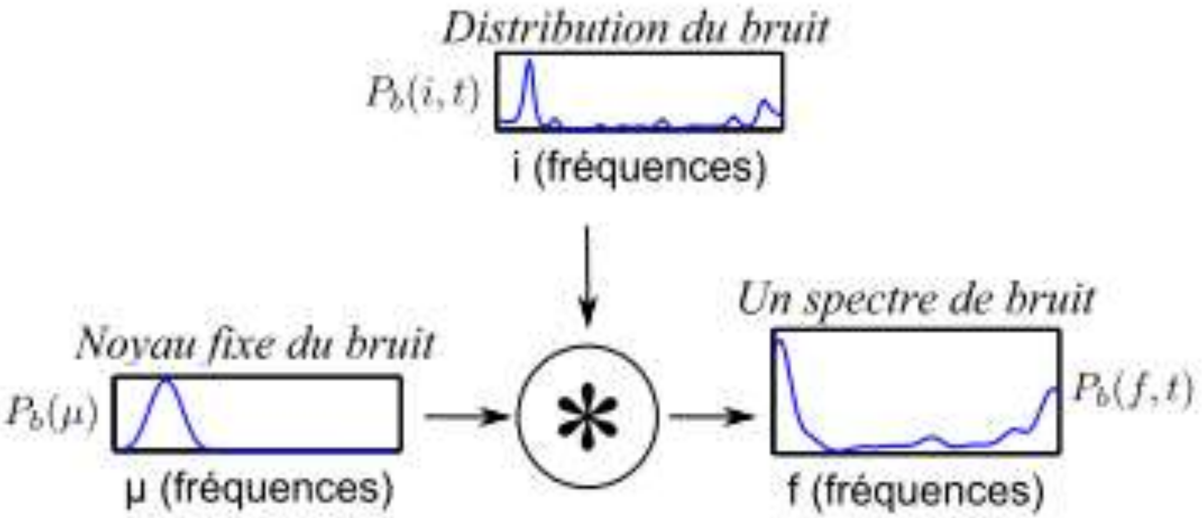


B. Fuentes, R. Badeau, and G. Richard, "Harmonic Adaptive Latent Component Analysis of Audio and Application to Music Transcription" IEEE Trans. On ASLP, 2013.

A PLCA model example

- **The HALCA model (Fuentes & al.)**

$$P(f,t) = P(c=h)P_h(f,t) + P(c=b)\boxed{P_b(f,t)}$$

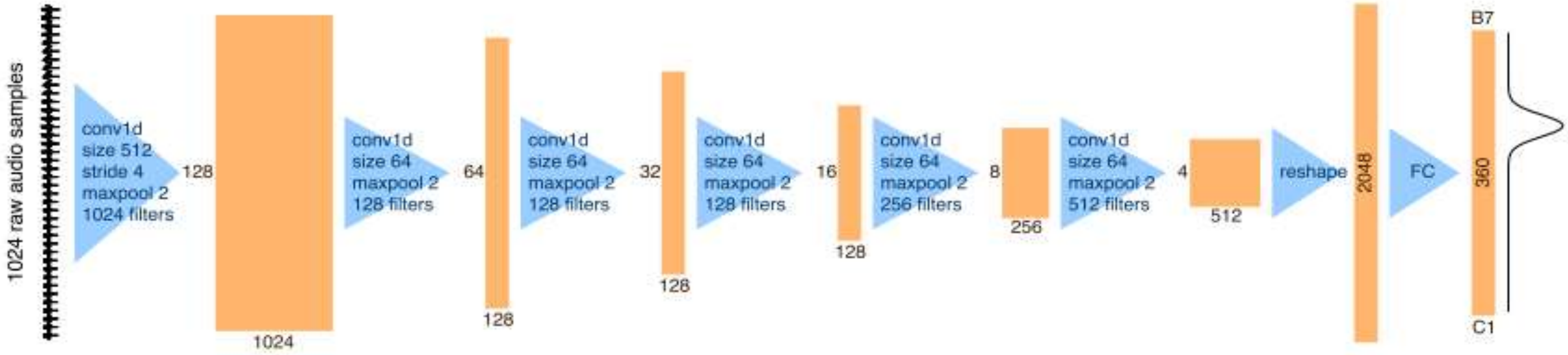


Pitch and multipitch estimation using DNN

- Example of Pyin as state of the art extension of YIN
- For multipitch
  - Cuesta (simple method targetted for homogeneous sources such as singing voices) H. Cuesta, B. McFee, and E. G´omez, “Multiple f0 estimation in vocal ensembles using convolutional neural networks,” in Proc. ISMIR, 2020,
  - pp. 302–309.

CRÉPE: A deep learning model for monopitch estimation (1/2)

- **Exploiting Machine learning (deep learning) for pitch estimation**



- **Output:**
  - 360 nodes (20 cents apart (1/5th of a semitone) from C1 ou B7) $\mathrm{¢}(f) = 1200 \cdot \log_2 \frac{f}{f_{\text{ref}}}$
  - Pitch estimate is the weighted mean of the output:

$$\hat{\mathrm{¢}} = \frac{\sum_{i=1}^{360} \hat{y}_i \mathrm{¢}_i}{\sum_{i=1}^{360} \hat{y}_i},$$

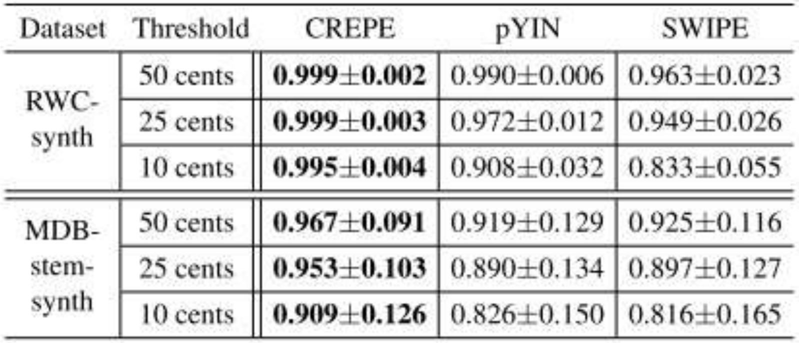
  - Trained with binary cross entropy loss

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^{360} \left(-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)\right) \qquad y, \hat{y} \in \mathbb{R}_{[0-1]}$$

Kim, Jong Wook et al. “Crepe: A Convolutional Representation for Pitch Estimation.” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018): 161-165.

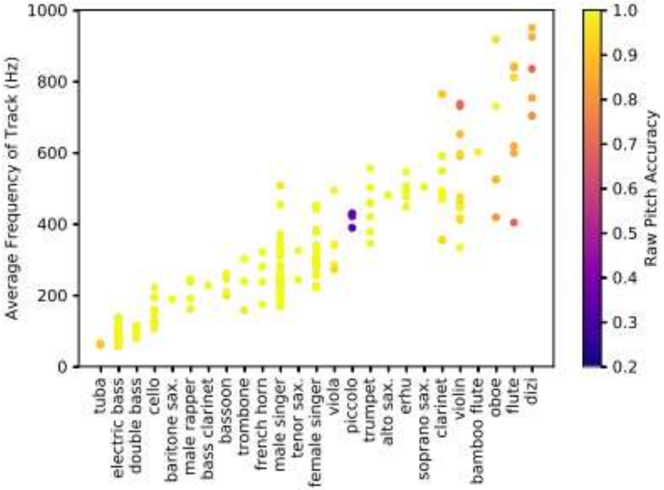
CREPE: A deep learning model for monopitch estimation (2/2)

- **A few results**



Dataset	Threshold	CREPE	pYIN	SWIPE
RWC-synth	50 cents	**0.999±0.002**	0.990±0.006	0.963±0.023
	25 cents	**0.999±0.003**	0.972±0.012	0.949±0.026
	10 cents	**0.995±0.004**	0.908±0.032	0.833±0.055
MDB-stem-synth	50 cents	**0.967±0.091**	0.919±0.129	0.925±0.116
	25 cents	**0.953±0.103**	0.890±0.134	0.897±0.127
	10 cents	**0.909±0.126**	0.826±0.150	0.816±0.165

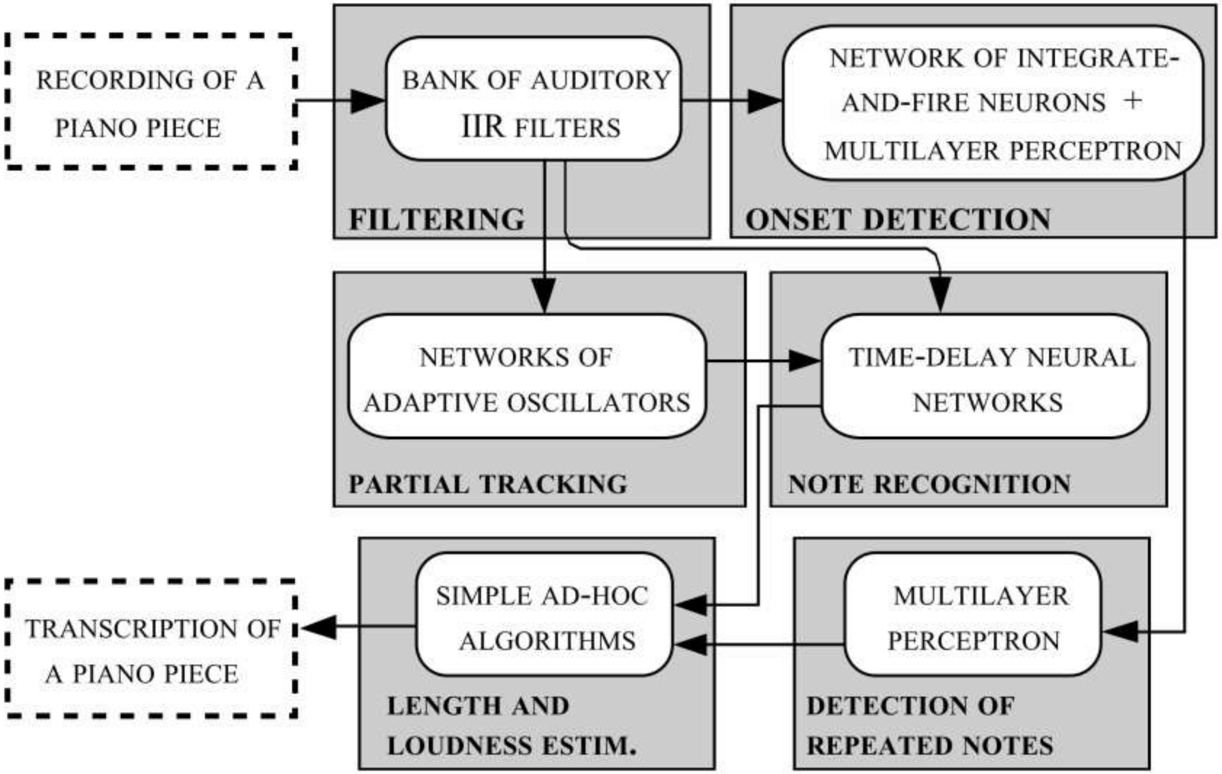
- **Better performances for low frequencies***



**: some errors due small Numbers of sound exemples for some instruments*

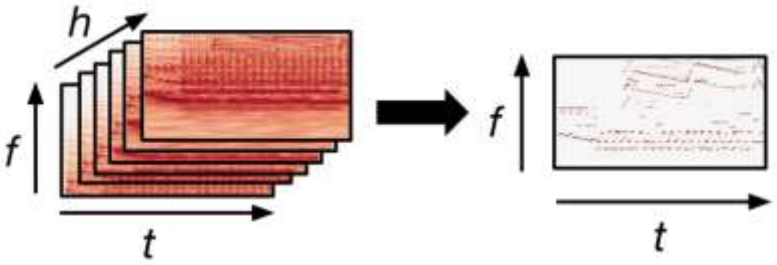
Multipitch estimation using neural networks

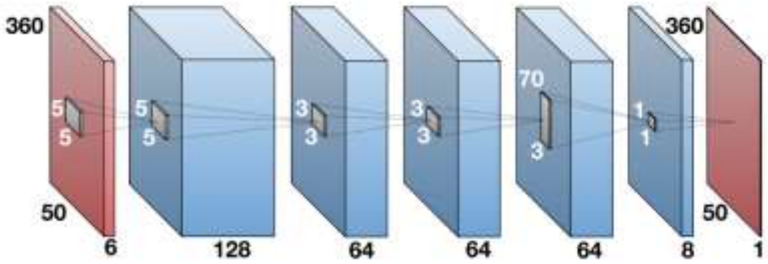
- **An early example by M. Marolt (2004) for piano sounds**


RECORDING OF A PIANO PIECE
BANK OF AUDITORY IIR FILTERS
FILTERING
NETWORK OF INTEGRATE-AND-FIRE NEURONS + MULTILAYER PERCEPTRON
ONSET DETECTION
NETWORKS OF ADAPTIVE OSCILLATORS
PARTIAL TRACKING
TIME-DELAY NEURAL NETWORKS
NOTE RECOGNITION
TRANSCRIPTION OF A PIANO PIECE
SIMPLE AD-HOC ALGORITHMS
LENGTH AND LOUDNESS ESTIM.
MULTILAYER PERCEPTRON
DETECTION OF REPEATED NOTES

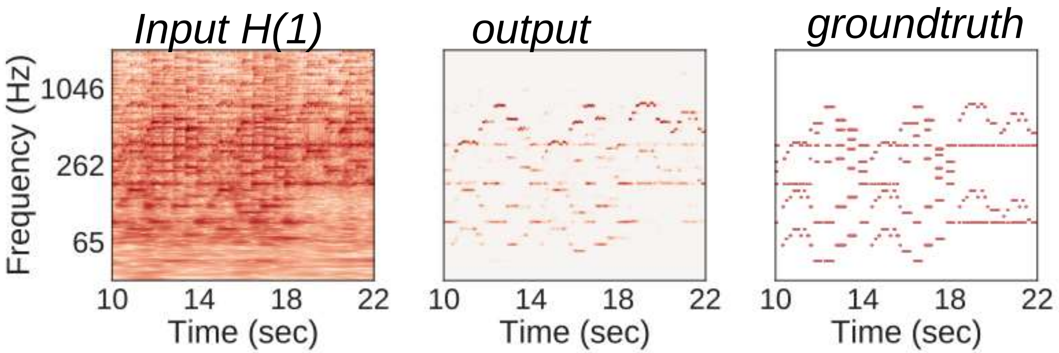
Marolt, Matija. (2004). A Connectionist Approach to Automatic Transcription of Polyphonic Piano Music. Multimedia, IEEE Transactions on. 6. 439 - 449. 10.1109/TMM.2004.827507.

Multipitch estimation using neural networks



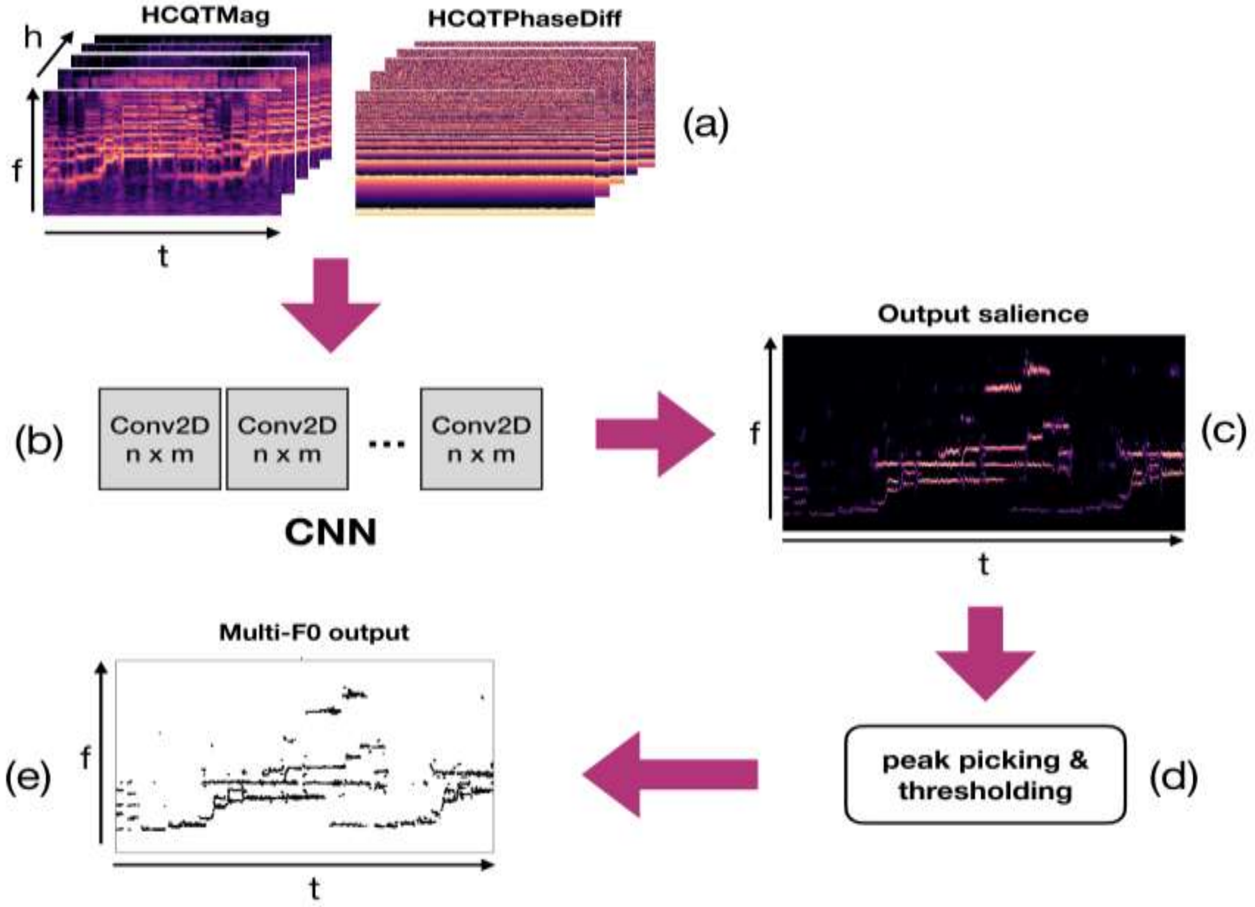


- Use of a specific input representation: the harmonic-CQT $f_k = h \cdot f_{\min} \cdot 2^{k/B}$
- CNN architecture with Relu ; Last layer with sigmoid
- The predicted saliency map can be interpreted as a likelihood score of each time-frequency bin belonging to an f0 contour.



Bittner, Rachel & McFee, Brian & Salamon, Justin & Li, Peter & Bello, Juan. (2017). Deep Salience Representations for f0 Estimation in Polyphonic Music. In proc ISMIR 2017

An extension focus on singing voices



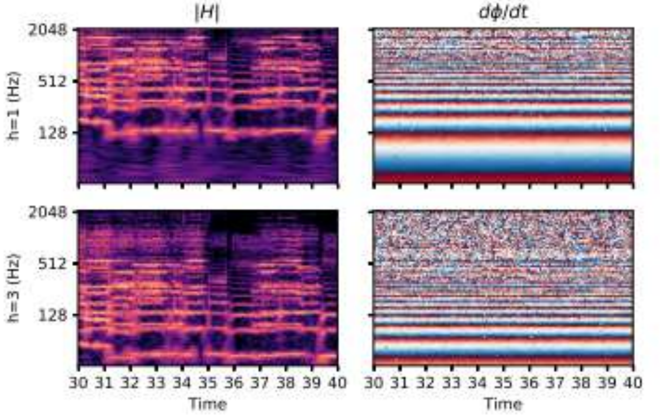
H. Cuesta, B. McFee, and E. Gomez, “Multiple f0 estimation in vocal ensembles using convolutional neural networks,” in Proc. ISMIR, 2020,

An extension focus on singing voices

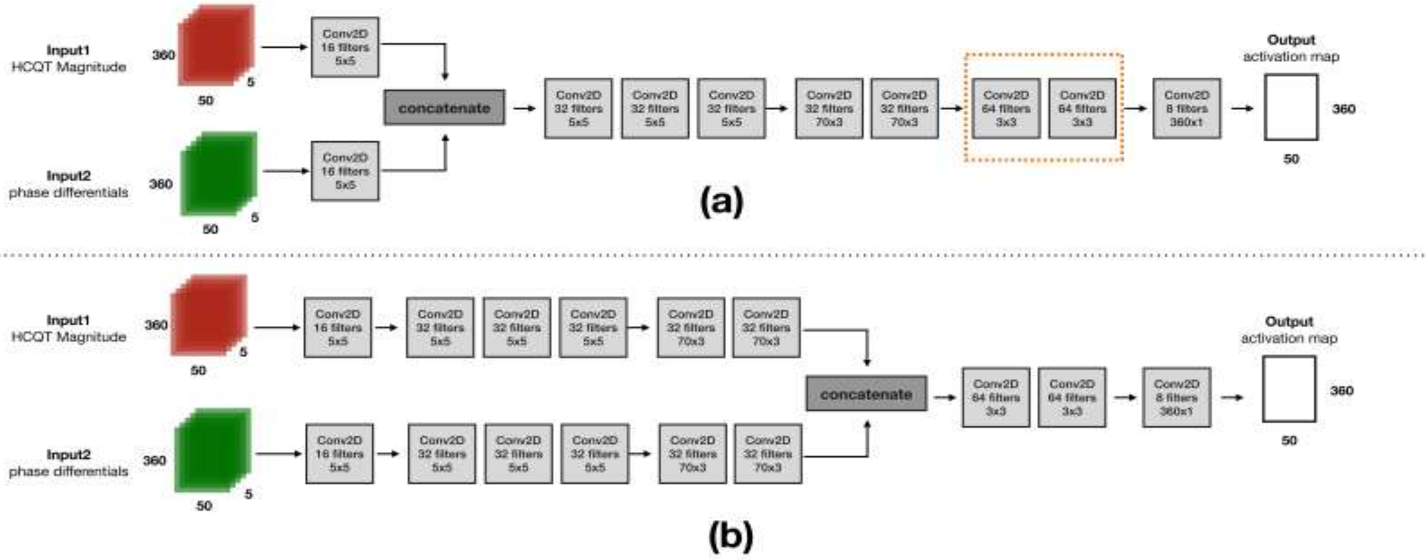
- Extended input features with HCQT Phase

(phase is directly linked to Instantaneous frequency)

$$\omega_{ins} = \frac{\delta\phi(t)}{\delta t} \rightarrow f_{ins} = \frac{1}{2\pi}\frac{\delta\phi(t)}{\delta t}$$

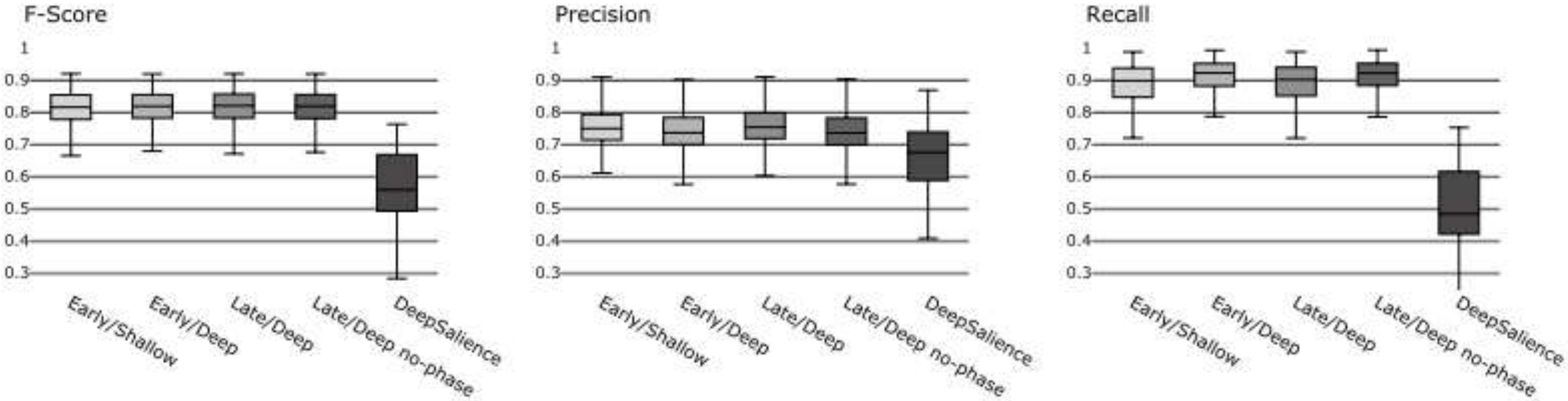


- New architectures (with fusion of input)



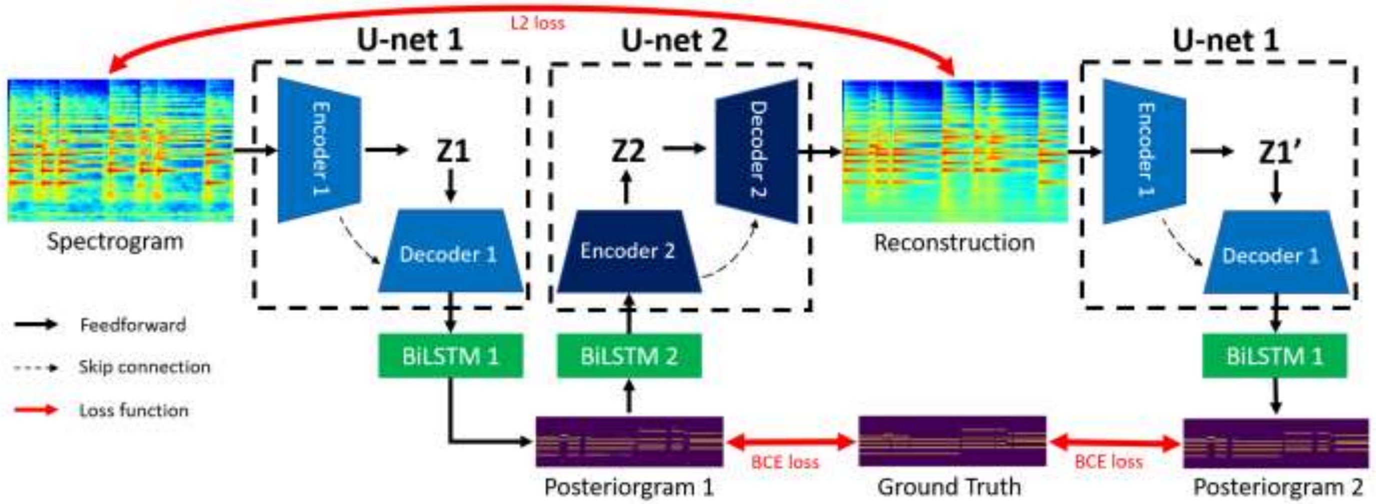
An extension focus on singing voices

- An idea of the performances (test sets > 3000 audio files)



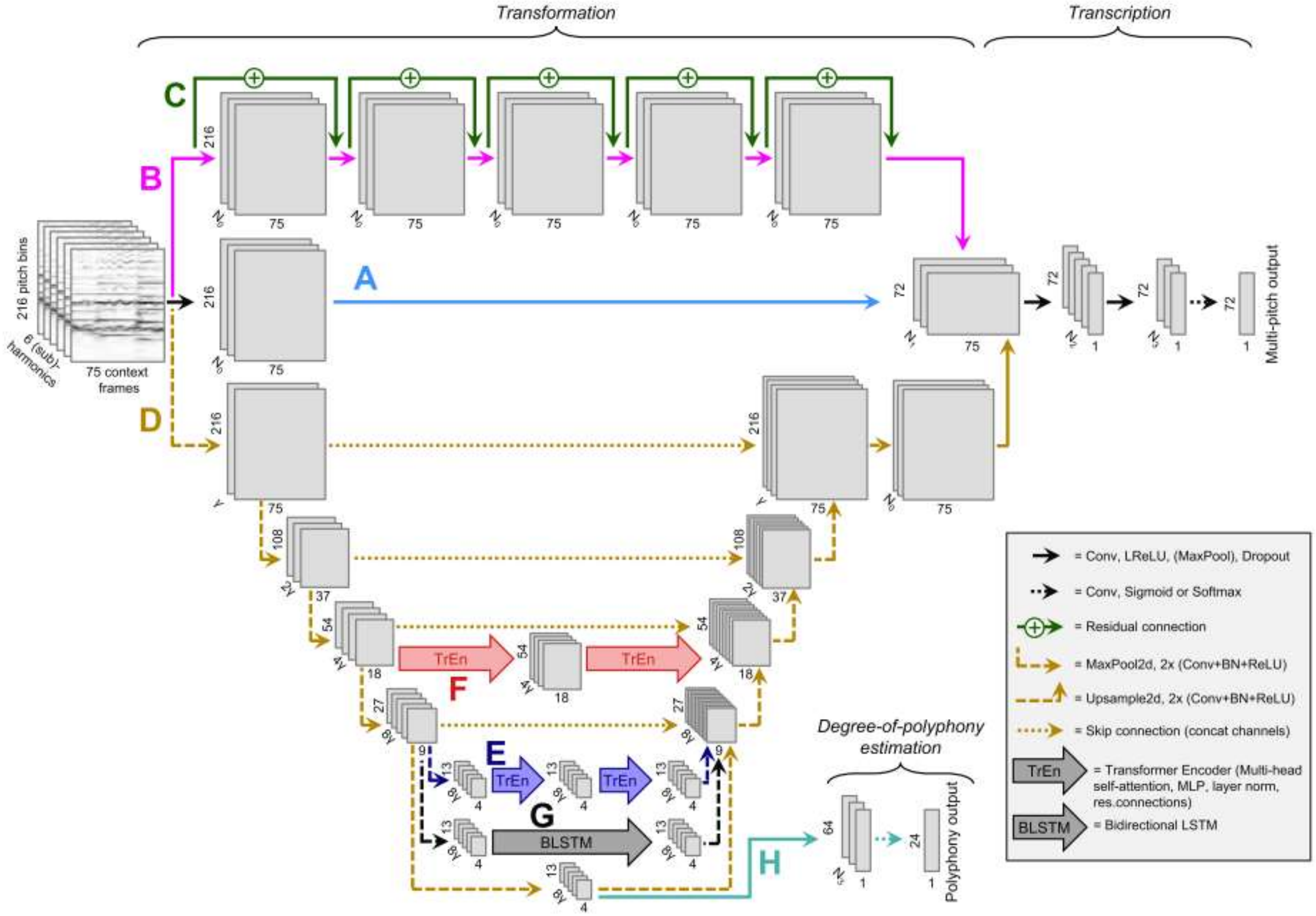
Multipith estimation using Unets (with spectrogram reconstruction)

- **Intuition: we mimic the human behaviour when evaluating a transcription:**
  - We « listen » to the transcription
  - We optimise the algorithm to reduce the errors



Cheuk, Kin Wai et al. "The Effect of Spectrogram Reconstruction on Automatic Music Transcription: An Alternative Approach to Improve Transcription Accuracy." 2020 25th International Conference on Pattern Recognition (ICPR) (2020): 9091-9098.

U-net architectures for multipitch estimation


Transformation
Transcription
C
B
A
D
E
F
G
H
216 pitch bins
6 (sub)-harmonics
75 context frames
216
75
N_0
72
N_1
N_2
N_3
1
Multi-pitch output
108
37
54
18
27
9
13
4
γ
2γ
4γ
8γ
TrEn
BLSTM
Degree-of-polyphony estimation
64
24
Polyphony output
= Conv, LReLU, (MaxPool), Dropout
= Conv, Sigmoid or Softmax
= Residual connection
= MaxPool2d, 2x (Conv+BN+ReLU)
= Upsample2d, 2x (Conv+BN+ReLU)
= Skip connection (concat channels)
= Transformer Encoder (Multi-head self-attention, MLP, layer norm, res.connections)
= Bidirectional LSTM

C. Weiß and G. Peeters, "Comparing Deep Models and Evaluation Strategies for Multi-Pitch Estimation in Music Recordings," in *IEEE/ACM Trans. On AASP*, vol. 30, pp. 2814-2827, 2022, doi: 10.1109/TASLP.2022.3200547

Multipitch estimation using neural networks: other neural approaches

- Deep spiking networks [5]
- Multi-resolution spectrogram as input with LSTM networks [4]
- Use of a kind of “language model” in Neural Autoregressive Distribution Estimator, also known as NADE (*similar to wavenet architecture*) [3]
- A succession of 2 bi-LSTM networks (for note onset detection and note duration estimation), in [2]
- Unet networks (with self-attention [6], spectrogram reconstruction [7], varied architectures [8])

- An interesting reading: [1]

« Yet, despite these [...] limitations, NMF-based methods remain competitive or even exceed the results achieved using NNs.”

[1] E. Benetos, S. Dixon, Z. Duan and S. Ewert, "Automatic Music Transcription: An Overview," in *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20-30, Jan. 2019, doi: 10.1109/MSP.2018.2869928.

[2] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. S. C. Raffel, J. Engel, S. Oore, and D. Eck, “Onsets and frames: Dual-objective piano transcription,” in Proc. Int. Society Music Information Retrieval Conf., 2018, pp. 50–57.

[3] S. Sigtia, E. Benetos, and S. Dixon, “An end-to-end neural network for polyphonic piano music transcription,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 5, pp. 927–939, 2016.

[4] S. Böck and M. Schedl, “Polyphonic piano note transcription with recurrent neural networks,” in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2012, pp. 121–124.

[5] Qian, Hanxiao et al. “Robust Multipitch Estimation of Piano Sounds Using Deep Spiking Neural Networks.” 2019 IEEE Symposium Series on Computational Intelligence (SSCI) (2019): 2335-2341.

[6]Y. -T. Wu, B. Chen and L. Su, "Multi-Instrument Automatic Music Transcription With Self-Attention-Based Instance Segmentation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2796-2809, 2020, doi:

[8] C. Weiß and G. Peeters, "Comparing Deep Models and Evaluation Strategies for Multi-Pitch Estimation in Music Recordings," in *IEEE/ACM Trans. On AASP*, vol. 30, pp. 2814-2827, 2022, doi: 10.1109/TASLP.2022.3200547.