A PITCH SALIENCE FUNCTION DERIVED FROM HARMONIC FREQUENCY DEVIATIONS FOR POLYPHONIC MUSIC ANALYSIS

A. Degani, R. Leonardi, P. Migliorati
University of Brescia
DII, Signals and Communication Lab
38, Via Branze - 25123 Brescia (ITALY)
a.degani@unibs.it

*G. Peeters**
STMS - IRCAM - CNRS - UPMC
Sound Analysis and Synthesis
1,pl. Igor Stravinsky - 75004 Paris (FRANCE)
peeters@ircam.fr

ABSTRACT

In this paper, a novel approach for the computation of a pitch salience function is presented. The aim of a pitch (considered here as synonym for fundamental frequency) salience function is to estimate the relevance of the most salient musical pitches that are present in a certain audio excerpt. Such a function is used in numerous Music Information Retrieval (MIR) tasks such as pitch, multiple-pitch estimation, melody extraction and audio features computation (such as chroma or Pitch Class Profiles). In order to compute the salience of a pitch candidate f, the classical approach uses the weighted sum of the energy of the short time spectrum at its integer multiples frequencies hf. In the present work, we propose a different approach which does not rely on energy but only on frequency location. For this, we first estimate the peaks of the short time spectrum. From the frequency location of these peaks, we evaluate the likelihood that each peak is an harmonic of a given fundamental frequency. The specificity of our method is to use as likelihood the deviation of the harmonic frequency locations from the pitch locations of the equal tempered scale. This is used to create a theoretical sequence of deviations which is then compared to an observed one. The proposed method is then evaluated for a task of multiple-pitch estimation using the MAPS test-set.

1. INTRODUCTION

A salience function is a function that provides an estimation of the predominance of different frequencies in an audio signal at every time frame. It allows to obtain an improved spectral representation in which the fundamental frequencies have a greater relevance compared to the higher partials of a complex tone. The computation of a salience function is commonly used as a first step in melody, predominant-pitch (pitch is considered here as synonym to fundamental frequency or f_0) or multiple-pitch estimation systems [1, 2, 3, 4].

1.1. Classical approach

In the classical approach [5], the salience (or strength) of each f_0 candidate is calculated as a weighted sum of the amplitudes of the spectrum at its harmonic frequencies (integer multiples of f_0). In the discrete frequency case, this can be express as:

$$S[k] = \sum_{h=1}^{H} w_h |X[hk]| \quad (1)$$

where k is the spectral bin, H is the number of considered partials, w_h is a partials' weighting scheme and $|X[k]|$ is the amplitude spectrum. This process is repeated for each time frame m. In this approach, the choice of the number of considered harmonics H and the used weighting scheme w_h are important factors and directly affect the obtained results [5]. The weighting scheme w_h implicitly models the sound source. Since the classical approach is based on the amplitude/energy of the spectrum, it is sensitive to the timbre of the sources. In order to make the estimation more robust against timbre variations, spectral whitening or flattening processes have been proposed [2, 6, 7, 8].

Among other approaches, the one of [9] proposes to estimate the salient pitch of a complex tone mixture using a psychoacoustic motivated approach. It uses the notions of masking and virtual pitch (sub-harmonic coincidence) calculation.

1.2. Proposal

In this paper, we propose a novel salience function which does not rely on the amplitude/energy of the spectrum but only on the frequency location of the peaks of the spectrum. Doing this, our method is not sensitive to timbre variations hence does not necessitate whitening processes.

The specificity of our method is to use as likelihood the deviation of the harmonic frequency locations from the pitch locations of the equal tempered scale. This is illustrated in Figure 1, for the harmonic frequencies of the pitch $C4$ (MIDI Key Number $i = 60$), which 3-rd and 6-th harmonic frequencies are slightly above the pitches $i = 79$ and $i = 81$ respectively, while its 5-th and 7-th are below the pitches $i = 88$ and $i = 94$ respectively (in the equal tempered scale). This is used to create a theoretical sequence of deviations which is then compared to an observed one derived from the peaks detected in the spectrum.

Paper organization: In section 2 we present the motivation behind the concept of this novel salience computation approach (section 2.2) and the details of its computation (sections 2.3, 2.4, 2.5 and 2.6). In section 3 we propose a basic evaluation framework of salience function based on a multi-pitch estimation paradigm (section 3.1) and asses the performances of our proposed method (section 3.4). We finally conclude in section 4 and provide directions for future works.

* Thanks to the Quaero Program funded by Oseo French State agency for innovation.

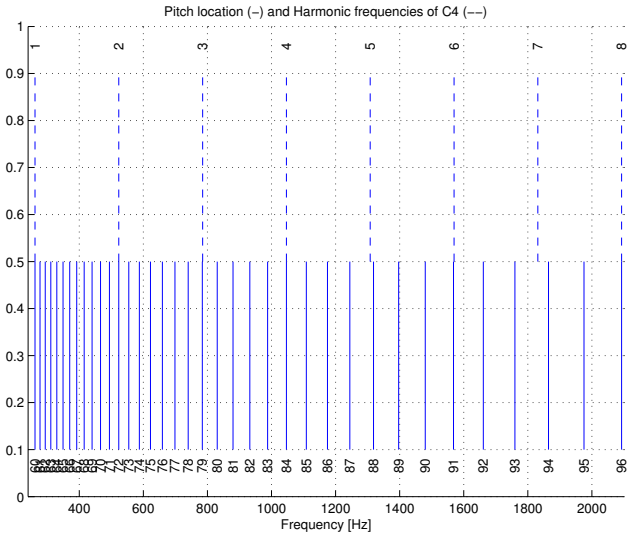

Figure 1: (Lower part) Frequency location of the pitches of the equal tempered scale for a tuning of 440 Hz. (Upper part) Frequencies of the harmonic series of the pitch C4 (261.6 Hz).

2. PROPOSED METHOD

2.1. Overview

The global flowchart of our method is represented in Fig. 2.

The content of the audio signal is first analyzed using Short Time Fourier Transform (STFT). At each time frame m, the peaks of the local Discrete Fourier Transform (DFT) are estimated using a peak-picking algorithm.
We denote by $\mathcal{P}_m = \{(f_1, a_1), \ldots, (f_P, a_P)\}$ the set of peaks detected at the frame m where f_p and a_p are the frequency and amplitude of the p-th peak. Since our salience function is based on an equal tempered cents grid, we then need to estimate the tuning frequency f_{ref} of the audio signal. We then compute at each frame m the salience value of each peak p by comparing its frequency to the ones of an equal tempered scale tuned on f_{ref}. This salience allows to discriminate peaks which are fundamental frequency from the ones that are harmonic partials.

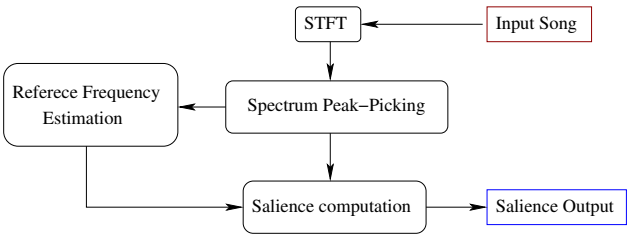

Figure 2: General scheme of the method.

2.2. Motivations for using frequency deviations for pitch salience computation

The computation of our salience function only relies on the frequency positions of the peaks of the spectrum (not on their energy). The basic idea we develop is the following: for a given note at fundamental frequency f_0 its h-th harmonic frequency exhibits a specific deviation from the equal tempered scale. For example, for a tuning at 440 Hz, the third ($h = 3$) harmonic of a A4 note ($f_0 = 440$ Hz) is at frequency 1320 Hz while the closest note of the equal tempered scale is at 1318.5 Hz. The specific deviation of the third harmonic is then 1.95 cents.

For a given frequency f_0, the frequency of its h-th harmonic is defined by

$$f_h^{f_0} = h \cdot f_0 \tag{2}$$

The deviation in cents of the harmonic $f_h^{f_0}$ from the equal tempered grid is defined as:

$$d_h^{f_0} = 100 \left[12 \log_2 \left(\frac{f_h^{f_0}}{f_{ref}} \right) - \left\lfloor 12 \log_2 \left(\frac{f_h^{f_0}}{f_{ref}} \right) \right\rceil \right] \tag{3}$$

where $\lfloor \cdot \rceil$ is the rounding operator and f_{ref} is the $A4$ tuning frequency estimated from the data[1]. We denote by $\{f_h^{f_0}\}$, the sequence of all the harmonic frequencies of f_0 and by $\{d_h^{f_0}\}$, the theoretical sequence of deviations.

This deviation is independent[2] of the actual f_0. We therefore simply denote it by $\{d_h\}$ in the following. In Fig. 3, we illustrate the deviation of the first 20 harmonics of a complex tone from the equal tempered note scale.

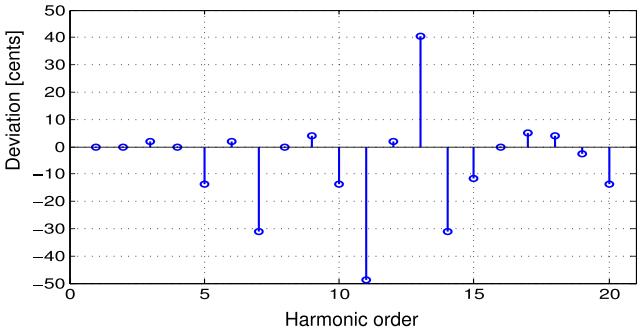

Figure 3: Deviation of the first 20 harmonic frequencies of a complex tone from the pitch of the equal tempered scale.

Salience computation: Since the sequence $\{d_h\}$ is independent of fundamental frequency, we can simply compute the salience of each f_0 candidate at frequency f_p, as the correlation between

[1] Or blindly chosen as 440 Hz.

[2] **Proof that $d_h^{f_0}$ is independent of f_0:** Under the hypothesis that the analysed musical excerpt is played on the equal temperament scale and using an accurately tuned instrument, we can calculate each equal-tempered note frequency f_i in the audio spectrum using an integer number i as follows:

$$f_i = f_{ref} \cdot 2^{\left(\frac{i}{12}\right)} \tag{4}$$

The integer number i represents the note index in the MIDI notation without the offset of 69 (for the sake of simplicity, we assume that $A4$ correspond to $i = 0$ instead of $i = 69$). Now, it is easy to check that for all fundamental frequencies $f_0 = f_i$, (3) can be rewritten as:

$$\begin{aligned} d_h^{f_i} &= 100 \left[12 \log_2(h) + i - \lfloor 12 \log_2(h) + i \rceil \right] \\ &= 100 \left[12 \log_2(h) - \lfloor 12 \log_2(h) \rceil \right] \end{aligned} \tag{5}$$

Since $i \in \mathbb{Z}$, we can say that $\lfloor 12 \log_2(h) + i \rceil = i + \lfloor 12 \log_2(h) \rceil$ and it is clear that the sequence $\{d_h^{f_i}\}$ does not depend on the fundamental frequency f_i.

the theoretical sequence of deviation $\{d_h\}$ and the measured sequence of the deviation $\{\hat{d}_h^{f_p}\}$. The measured sequence of deviations is the one corresponding to the peak detected in the spectrum $\mathcal{P}_m$.

Extension to inharmonic signals: Inharmonicity is a phenomenon related to the physical characteristics of a non-ideal string. The frequencies of the modes of vibration of an ideal string are exact integer multiples of the fundamental, but the stiffness of the material of the real strings shifts the modes of vibration at non-integer multiples [10].
In mathematical terms, the relation between the h-th partial $f_h^{f_0}$ and the fundamental frequency f_0 can be modelled as

$$f_h^{f_0}(\beta) = h f_0 \sqrt{1+\beta h^2} \tag{6}$$

where β is the *inharmonicity coefficient* which is related to the physical properties of a string. In order to take into account inharmonicity we use (6) instead of (2) into equation (3). It should be noted that whatever inharmonicity is taken into account or not, the theoretical sequence of deviations is always independent of f_0. However, the theoretical sequence of deviations now depends on the parameter β and it is denoted by $\{d_h(\beta)\}$.

In the next sections, we describe in details each block of our algorithm (see Fig. 2).

2.3. Short Time Fourier Transform

The N-terms STFT, at time frame m, of a discrete signal $x[n]$ is defined as

$$X_{m,k} = \sum_{n=0}^{N-1} x[n+\tau m] \cdot \bar{w}[n] \cdot e^{-j2\pi \frac{k}{N} n} \tag{7}$$

where $k \in [-N/2+1, \ldots, N/2]$, τ is the hop size (in samples) from two subsequent frames and $\bar{w}[n]$ is the windowing function. For our computation, we only use the amplitude of the STFT denoted by $|X_{m,k}|$. We use $N = 4096$ samples (which corresponds to 92.9 ms for a sampling rate of 44.1 KHz), $\tau = 2048$ samples (overlap of 50%) and $\bar{w}[n]$ is a Hanning windowing function.

2.4. Spectrum Peak Picking

In order to detect the local peaks of the spectrum, we use the algorithm proposed in the context of the Sinusoidal Modelling Synthesis framework (SMS) [11, 12]. In this context, a fixed number P of local maxima is detected in the amplitude spectrum $|X_{m,k}|$. For each local maximum, its frequency $\mathcal{M}_p$ is refined using a 3-point parabolic interpolation using $[\mathcal{M}_p - 1, \mathcal{M}_p, \mathcal{M}_p + 1]$. The obtained frequency is denoted by f_p in Hz. The result of the peak picking algorithm is the sequence $\mathcal{P}_m = \{(f_1, a_1), \ldots, (f_P, a_P)\}$ made of pairs of peaks frequency location f_p and amplitude a_p. The peak-picking is performed at each time frame $m \in [1 \ldots M]$. The concatenation of all peaks sequences, $\mathcal{P}_{tot} = \mathcal{P}_1 \| \mathcal{P}_2 \| \ldots \| \mathcal{P}_M$, is used as input for the reference tuning estimation algorithm.

2.5. Reference Frequency Estimation

Since our algorithm relies on the equal-tempered cent scale, the tuning f_{ref} (or reference frequency) of the audio signal being analyzed need to be estimated. For this, we use the method presented in [13]. This approach is entirely based on the observation that the deviation d is a periodic measure and not an absolute measure, since it is a "wrapped around" quantity that should be evaluated from the nearest 100 cents grid point. Each cent value is mapped onto a unit circle 100 cents-periodic, and represented as a vector as follows

$$\mathbf{u}_p = a_p \cdot e^{j\phi_p} \tag{8}$$

where

$$\phi_p = \frac{2\pi}{100} \cdot 1200 \cdot \log_2\left(\frac{f_p}{440}\right) \tag{9}$$

and a_p is the peak amplitude used to weight each vector to avoid high impact of small (noise) peaks. We take the mean vector $\hat{\mathbf{u}}$ of all circular quantities $\mathbf{u}_p$ as follows

$$\hat{\mathbf{u}} = \frac{\sum_{p=1}^{P_{tot}} \mathbf{u}_p}{\sum_{p=1}^{P_{tot}} a_p} \tag{10}$$

where P_{tot} is the element count of the concatenated sequence $\mathcal{P}_{tot}$. The overall deviation is then computed from the angle of the resulting vector $\hat{\mathbf{u}}$, that is

$$D = \frac{1}{2\pi} \angle(\hat{\mathbf{u}}) \tag{11}$$

The reference frequency of the entire music piece can be computed as

$$f_{ref} = 440 \cdot 2^{\frac{D}{12}} \tag{12}$$

2.6. Salience Function Computation

As previously said, the salience $S_p(\beta)$ of a given peak p can be calculated as the correlation C between the theoretical sequence of deviation $\{d_h(\beta)\}$ and the measured one $\{\hat{d}_h^{f_p}(\beta)\}$. From the abstract point of view, $S_p(\beta)$ is calculated using:

$$S_p(\beta) = C\left(\{d_h(\beta)\}, \{\hat{d}_h^{f_p}(\beta)\}\right) \tag{13}$$

where $C(\cdot,\cdot)$ is a generic correlation measure. The two deviation sequences can be seen as two vectors $\mathbf{d}(\beta) = [d_1(\beta), \ldots, d_H(\beta)]$ and $\hat{\mathbf{d}}^p(\beta) = [\hat{d}_1^{f_p}(\beta), \ldots, \hat{d}_H^{f_p}(\beta)]$, so that, a good correlation measure can be the inner product $< \cdot, \cdot >$. In practice, in order to reduce the influence of very small values (hence often noisy) in the computation of the salience, the correlation is weighted by the local amplitude a_p of the f_0 candidate f_p:

$$S_p(\beta) = a_p < \mathbf{d}(\beta), \hat{\mathbf{d}}^p(\beta) > = a_p \sum_{h=1}^{H} d_h(\beta) \cdot \hat{d}_h^{f_p}(\beta) \tag{14}$$

Computation of $\hat{d}_h^{f_p}(\beta)$: $\{f_h^{f_p}(\beta)\}$ is the sequence made of the harmonic frequencies of a detected peak p: $f_h^{f_p}(\beta) = h f_p \sqrt{1+\beta h^2}$. $\{\hat{d}_h^{f_p}(\beta)\}$ is the vector of measured deviations corresponding to $\{f_h^{f_p}(\beta)\}$. $\{\hat{d}_h^{f_p}(\beta)\}$ is computed for all the detected peaks $p \in \mathcal{P}_m$ at frame m (i.e. we consider each detected peak as a potential pitch candidate).

To validate a given pitch candidate f_p, we look among the detected peaks the ones that are harmonics of this candidate. This is done by using a function G centered on the $h-$th harmonic of f_p and evaluated at the detected peaks $f_{p'}$.

More precisely, $G(f_{p'}; \mu_{h,p}(\beta), \sigma_{h,p}(\beta))$ is a Gaussian function evaluated at $f_{p'}$, with

- mean $\mu_{h,p}(\beta) = h f_p \sqrt{1+\beta h^2}$ and

- standard deviation $\sigma_{h,p}(\beta) = \mu_{h,p}(\beta)\left(1 - 2^{\frac{\alpha}{1200}}\right)$

where the parameter $\alpha = 20$ cents is chosen experimentally in order to take into account the effect of the frequency location error of the peak picking step. We chose the α such that the number of the False Positive is reduced without losing Precision (see the section 3.2 for the explanation of the evaluation measures).
The Gaussian function we use, has a maximum value of one when $f_{p'} = \mu_{h,p}(\beta) = hf_p\sqrt{1+\beta h^2}$; in other words G will only take non-zero values for the $f_{p'}$ (the detected peaks) which are close to $hf_p\sqrt{1+\beta h^2}$.

To each detected peaks $f_{p'}$ is associated a deviation $\bar{d}_{p'}$ as defined in (3).

$$\bar{d}_{p'} = 100\left[12\log_2\left(\frac{f_{p'}}{f_{ref}}\right) - \left\lfloor 12\log_2\left(\frac{f_{p'}}{f_{ref}}\right)\right\rceil\right] \quad (15)$$

The deviation of $f_h^{f_p}(\beta)$ is then computed as the following weighted sum:

$$\hat{d}_h^{f_p}(\beta) = \sum_{p'=1}^{P} G(f_{p'}; \mu_{h,p}(\beta), \sigma_{h,p}(\beta)) \cdot \bar{d}_{p'} \quad (16)$$

A single value of β is assigned to each pitch candidate f_p. The typical range of β for a piano string [10] is $\beta \in B = \{0\} \cup [10^{-5}, 10^{-3}]$. In order to estimate β we maximize

$$S_p = \max_{\beta \in B}[S_p(\beta)] \quad (17)$$

Notice that in the practical case, all the values of β in the search range must be tested exhaustively because $S_p(\beta)$ is an "unpredictable" function and no numerical optimized algorithm can be used in order to find the maximum of that function. The maximization of (17) provides simultaneously the value of S_p and the one of the inharmonicity coefficient β for each spectral peak p. Of course, only the values of β corresponding to true notes make sense.

The limits of the equal temperament: Using the equal-tempered grid of semitones is fundamental for the consideration made in Sec. 2.2. Moreover, it is reasonable to think that only exact tuned instruments[3] are needed in order to maintain the validity of equation (5). However, the gaussian weighting scheme used in (16) ensures that the slighted deviated fundamental frequencies are not much negatively affected. However, the spectral peaks that are detuned more than $\pm\alpha$ cents can be excessively penalized.

3. EVALUATION

There is no standard method to evaluate the performances of a salience function by itself. This is because such a function is usually a pre-processing step of a more complicated algorithm (as for example a pitch-estimation method [2, 14]). Therefore, in order to be able to test our salience, we chose to construct a very simple and straightforward multiple-pitch estimation algorithm from our salience function. In section 3.1, we explain the post-processing applied to the salience function in order to obtain a multi-pitch estimation.

[3]For example, the octave stretching in piano tuning can be a problem.

3.1. Multiple-pitch estimation: post-processing of the salience function

In order to test our salience function as a multi-pitch estimation algorithm, we chose to apply a basic post-processing process that transforms the salience function into a piano-roll representation. The piano-roll $\hat{\mathcal{R}}_{m,i}$ can be seen as a spectrogram-like binary representation where the rows are the time frames m and the columns are the MIDI Key Number i[4] . If a note i is marked as detected at the time frame m, the corresponding element $\hat{\mathcal{R}}_{m,i}$ is set to 1; otherwise it is set to 0.

At each time frame m, we have a sequence of P pairs of peak frequency and salience values $\mathcal{S}_m = \{(f_1, S_1), \ldots, (f_P, S_P)\}$. We normalize the values S_p in order to obtain maximum amplitude of one at each time frame. The negative values of salience are set to zero. Each peak frequency f_p is quantized to the nearest MIDI Key Number using

$$i_p = 69 + 12\log_2\left(\frac{f_p}{f_{ref}}\right) \quad (18)$$

where 69 correspond to the MIDI Key Number associated to the note $A4$ in the MIDI Tuning Standard (MTS). In order to remove holes (estimation errors) in the middle of notes (disruption in the salience value), we then apply a sliding median filter of size L frames along the time dimension m. Finally the binary piano-roll is obtained by applying a fixed threshold T to the values of $\hat{\mathcal{R}}_{m,i}$. We set to 1 all the values that are above T, and 0 the other ones. In Fig. 4, an example of piano-roll transcription is shown and different colors are used in order to highlight the True Positive, False Positive and False Negative. Notice that a considerable number of False Positive are just after a True Positive in the same MIDI Key Number. This is caused by the release time of the piano sound that is extended by the reverberation time simulated in the recordings.

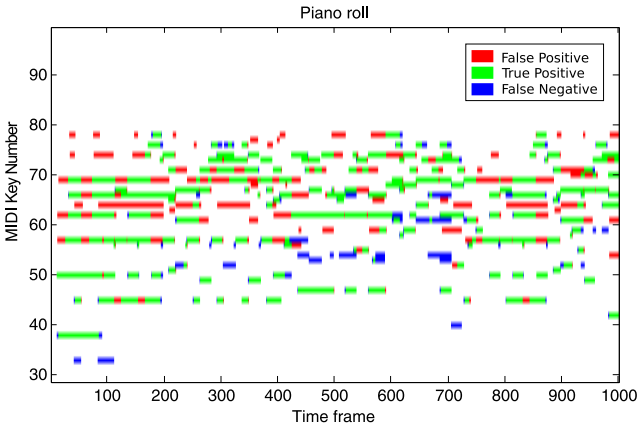

Figure 4: Piano roll representation $\hat{\mathcal{R}}_{m,i}$ obtained using our salience function. In this case $P = 0.65$, $R = 0.8$ and $F = 0.72$ (see explanation of the evaluation measures in section 3.2).

3.2. Evaluation measures

In order to evaluate our salience-based piano-roll, we have to compute a ground-truth piano-roll $\mathcal{R}_{m,i}$ for each song in the dataset.

[4]Ranging from 21 ($A0$) to 108 ($C8$).

$\mathcal{R}_{m,i}$ is obtained from the ground-truth text annotation that reports *onset* time, *offset* time and MIDI Key Number for each note played in a specific song. The note onset and offset time are quantized with the same hop size τ (converted in seconds) used by the algorithm.

We compare the ground-truth piano-roll $\mathcal{R}_{m,i}$ to the estimated piano-roll $\hat{\mathcal{R}}_{m,i}$ by comparing the values on each cell (m,i). We then compute the *Precision* (P), the *Recall* (R) and the *F-Measure* (F) defined as follows:

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad F = \frac{2PR}{P+R} \qquad (19)$$

where TP (True Positive) is the total number of correctly identified notes, FN (False Negative) of missed notes and FP (False Positive) the number of false notes detected.

3.3. Test-Set

Experiments are performed on the MIDI Aligned Piano Sounds test-set [15]. MAPS provides CD quality piano recording (44.1 kHz, 16-bit). This test-set is available under Creative Commons license and consists of about 40GB (65 hours) of audio files recorded using both real and synthesized pianos. The aligned ground-truth is provided as MIDI or plain text files. The alignment and the reliability of the ground-truth is guaranteed by the fact that the sound files are generated form this MIDI files with high quality samples or a Disklavier (real piano with MIDI input). In order to have a generalized test-set, the pianos have been played in different conditions, such as various ambient with different reverberation characteristics (9 combinations in total). This collection is subdivided into four different subsets. The set ISOL contains monophonic excerpts, MUS contains polyphonic music, UCHO is a set of usual chords in western music, and RAND is a collection of chords with random notes.

3.4. Results

Setting the parameters: The parameters of our algorithm are:

- H: the total number of considered harmonics,
- L: the length of the median filter,
- T: the salience threshold.

In order to tune these parameters we used the *AkPnStgb* audio files of the test-set[5]. The values that maximize the *F-Measure* are $H = 8$, $L = 6$ and $T = 0.2$. The total number of peaks per frame P, is not itself a parameter of the salience algorithm. $P = 40$ is chosen experimentally.

Harmonic vs Inharmonic model: We first compare in Figure 5 the Pitch estimation obtained by our model in harmonic setting (the β parameters is forced to 0) to the inharmonic setting (β is estimated). This is done using the whole MAPS test-set.

As we expected, taking into account the inharmonicity brings an improvement on overall. The precision P increases by 12% (from 0.43 to 0.55) and the F-Measure increases by 5%. Since the Recall does not change significantly, while the Precision does, we can say that considering string inharmonicity allows reducing the number of False Positive. Because the results are better with our inharmonic model, we only consider this one in the following.

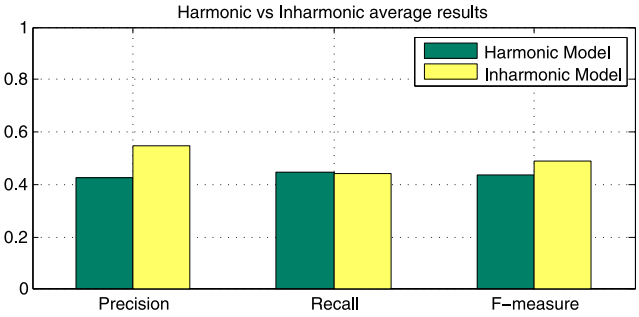

Figure 5: Pitch estimation results for our model in Harmonic (model forced to $\beta = 0$) vs Inharmonic setting (β is estimated).

Detailed Analysis: In Fig. 6, we provide the results in terms of Pitch estimation for each subset of MAPS using the inharmonic model. In Fig. 7, we provide the results in terms of Pitch-Class (i.e., without octave information). As we can see from the Figures, our approach is prone to octave errors. This is due to the fact that the deviation template itself does not exploit the octave information[6]. This octave ambiguity could only be solved with an ad hoc procedure. Figures 6 and 7 also show that on average, the precision P is greater than the recall R. For a fixed number of True Positive, this means that the number of False Negative (missed notes) is greater than the False Positive (added notes).

Influence of the T parameter: In Figure 8, we show the variation of the Recall and Precision in function of the choice of the parameter T (threshold on salience values). We see that the choice of T is a key parameter for the *Precision/Recall* trade-off, hence for the FP / FN trade-off. If our system is used as a front-end of a more complicated system which can filter-out the False Positives, we should use a value of T which maximizes Recall. It should be noted that Figure 8 is computed using only the MUS subset of MAPS. Because of this, the best value for T (in F-Measure sense) is $T = 0.1$ (which is different from the global optimum value for the entire MAPS test-set).

Comparison to state-of-the-art: In Table 1, we indicate the Pitch F-Measure results of our system in harmonic setting (P1, β is forced to 0) and in inharmonic setting (P2 , β is estimated). We compare our results to the ones obtained by Emiya et al. [15] and Benetos et al. [7] on the same test-set. Also, the results obtained by directly applying a threshold on the detected peaks are reported as a baseline results. As expected, the results obtained with our methods are not as good as the ones obtained with dedicated multi-pitch estimation algorithms.

The main reason is that our system is not a multi-pitch estimation method but only a pre-processing step to be used in a more complex system. Our straightforward post-processing procedure

[5] This is one of the nine different piano and recording condition set-up in the MAPS test-set

[6] In a hypothetical scenario where the peak peaking algorithm detects the peaks at frequency f_p and in an infinite number of its harmonics with amplitude equal to a_p, the salience value for the peak p and for the peaks with frequency $\bar{h} f_p$ with $\bar{h} = 2^j$, $j \in \mathbb{N}^+$ will be the same. To put it in another way, the peaks with frequency that is j octaves above f_p will measure the same salience value as f_p.

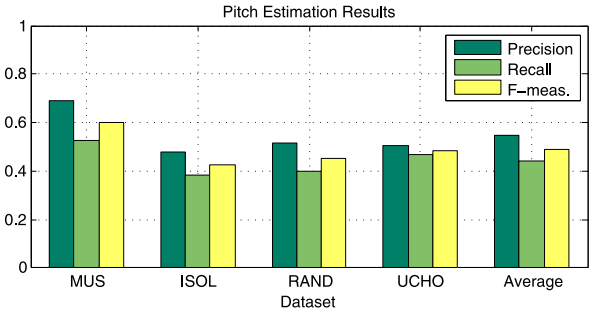

Figure 6: Pitch estimation result for each subset and the overall average (β is estimated).

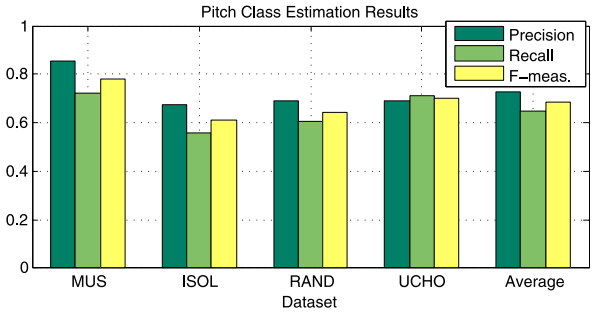

Figure 7: Pitch-Class estimation result for each subset and the overall average (β is estimated).

	Peak	P1	P2	Emiya et al.	Benetos et al.
F-Meas.	0.31	0.44	0.49	0.82	0.87

Table 1: Comparison of Pitch F-Measure results on MAPS test-set. *Peak* is a fixed threshold on detected peaks, *P1* is the proposed method without considering inharmonicity (β forced to 0) and *P2* is with the inharmonic model (β is estimated). *Emiya et al.* is presented in [15] and *Benetos et al.* in [7].

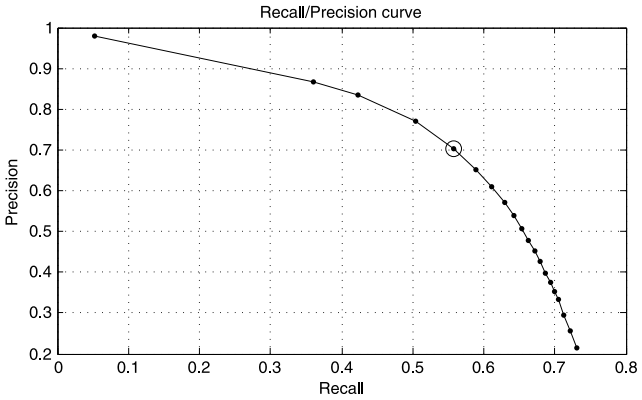

Figure 8: Pitch Recall/Precision curve for different values of T for the MUS subset. The best *F-Measure* (0.62) is obtained for $T = 0.1$ and is marked with the "O".

has been introduced only to asses the potential performances of our novel salience function design. In this context, our salience exhibit very promising results.

4. CONCLUSIONS

The performances obtained by our proposed salience function for the estimation of pitch-classes (Fig. 7) show that this kind of salience, even with simple post-processing procedure, is suitable for extracting audio features like Pitch Class Profile (PCP [16]) used in cover song detection or key/chord recognition tasks [17, 18]. Moreover, especially for a piano music test-set such as MAPS, considering the string inharmonicity is beneficial in terms of precision and F-Measure. Despite the fact that our salience function look promising, further development of an ad-hoc post-processing procedure is needed in order to be used for multi-pitch estimation. Moreover, as indicated in Section 3.4, the parameter T should be tuned depending on the application, in order to favour the F-Measure or the Recall. During our tests we have identified some weakness that are subjects for future research. The accuracy of the peak peaking algorithm is a key factor. A missing peak can negatively affect the overall accuracy performances. The octave ambiguity discussed in the previous section can be treated by developing specific procedure. Furthermore, the worst resolution in the low frequency spectrum can led to a large error when calculating the high order harmonic frequencies. Conversely, the note in the high portion of the audio spectrum does not have a sufficient number of partials to give a consistent value of salience because of the spectral roll-off near the Nyquist limit.

Acknowledgements

This work was partly founded by OSE through the Quaero project and by the French government Programme Investissements d'Avenir (PIA) through the Bee Music Project.

5. REFERENCES

[1] Karin Dressler, "Audio melody extraction for mirex 2009," *5th Music Information Retrieval Evaluation eXchange, (MIREX)*, 2009.

[2] Matti P. Ryynanen and Anssi P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, 2008.

[3] Yipeng Li and DeLiang Wang, "Pitch detection in polyphonic music using instrument tone models," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, vol. 2, pp. 481–484, 2007.

[4] Tiago Fernandes Tavares, Jayme Garcia Arnal Barbedo, and Amauri Lopes, "Improving a multiple pitch estimation method with ar models," *Proc. of Audio Engineering Society Conference: 42nd International Conference: Semantic Audio*, 2011.

[5] Justin Salamon, Emilia Gómez, and Jordi Bonada, "Sinusoid extraction and salience function design for predominant melody estimation," *Proc. of the 14th Int. Conference on Digital Audio Effects, (DAFx)*, 2011.

[6] Emmanouil Benetos and Simon Dixon, "Polyphonic music transcription using note onset and offset detection," *Proc.*

of IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 37–40, 2011.

[7] Emmanouil Benetos and Simon Dixon, "Multiple-f0 estimation of piano sounds exploiting spectral structure and temporal evolution," *Proc. of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, pp. 13–18, 2010.

[8] Justin Salamon and Emilia Gómez, "A chroma-based salience function for melody and bass line estimation from music audio signals," *Proc. of Sound and Music Computing Conference (SMC)*, pp. 331–336, 2009.

[9] Ernst Terhardt, Gerhard Stoll, and Manfred Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *Journal of the Acoustical Society of America*, vol. 71, pp. 679–688, 1982.

[10] Harvey Fletcher, E. Donnell Blackham, and Richard Stratton, "Quality of piano tones," *Journal of Acoustical Society of America*, vol. 34, no. 6, pp. 749–761, 1962.

[11] Xavier Amatriain, Jordi Bonada, Alex Loscos, and Xavier Serra, *in Udo Zölzer DAFX-Digital Audio Effects*, chapter Spectral processing, pp. 373–438, John Wiley & Sons, 2002.

[12] Xavier Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, A. Picialli C. Roads, S. Pope and G. De Poli, Eds., chapter Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122. Swets & Zeitlinger Publishers, 1997.

[13] Karin Dressler and Sebastian Streich, "Tuning frequency estimation using circular statistics," *Proc. of the 8th Int. Conf. on Music Information Retrieval, (ISMIR)*, pp. 357–360, 2007.

[14] Karin Dressler, "Pitch estimation by the pair-wise evaluation of spectral peaks," *Proc. of Audio Engineering Society Conference: 42nd International Conference: Semantic Audio*, 2011.

[15] Valentin Emiya, Roland Badeau, and Bertrand David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[16] Takuya Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music," *Proc. of the Int. Computer Music Conference, (ICMC)*, pp. 464–467, 1999.

[17] Joan Serrà, *Identification of Versions of the Same Musical Composition by Processing Audio Descriptions*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2011.

[18] Geoffroy Peeters, "Chroma-based estimation of musical key from audio-signal analysis," *Proc. of the 7th Int. Conf. on Music Information Retrieval, (ISMIR)*, 2006.