MULTIPITCH ESTIMATION OF QUASI-HARMONIC SOUNDS IN COLORED NOISE

Valentin Emiya, Roland Badeau, Bertrand David

GET - Télécom Paris (ENST), CNRS LTCI
46, rue Barrault, 75634 Paris cedex 13, France
valentin.emiya@enst.fr

ABSTRACT

This paper proposes a new multipitch estimator based on a likelihood maximization principle. For each tone, a sinusoidal model is assumed with a colored, Moving-Average, background noise and an autoregressive spectral envelope for the overtones. A monopitch estimator is derived following a Weighted Maximum Likelihood principle and leads to find the fundamental frequency (F_0) which jointly maximally flattens the noise spectrum and the sinusoidal spectrum. The multipitch estimator is obtained by extending the method for jointly estimating multiple F_0's. An application to piano tones is presented, which takes into account the inharmonicity of the overtone series for this instrument.

1. INTRODUCTION

Multipitch estimation is a critical topic for many applications, both in the field of speech processing (*e.g.* prosody analysis) [1] and in the context of musical signal analysis (*e.g.* automatic transcription) [2, 3]. The challenge offered by the spectral interference of the overtones of simultaneous notes has been taken up by various methods, some aiming at detecting a periodicity in the signal [4] or in its spectrum [5] while others use a combination of both spectral and temporal cues [6, 7]. Recent trends in the task include estimation in a bayesian framework [8] or in a perceptually compliant context [7]. The technique proposed in this paper is based on a Weighted Maximum Likelihood (WML) principle and belongs to the spectral estimators category.

This paper is organized as follows. Section 2 introduces the Maximum Likelihood principle applied to the proposed signal model. Section 3 then details the adaptation of the theoretical method to the multipitch estimation task in the case of piano sounds. Experimental results are given in section 4. Finally, conclusions are presented in section 5.

The research leading to this paper was supported by the French GIP ANR under contract ANR-06-JCJC-0027-01, Décomposition en Éléments Sonores et Applications Musicales - DESAM, and by the French Ministry of Education and Research under the Music Discover project of the ACI-Masse de données

2. WEIGHTED MAXIMUM LIKELIHOOD PITCH ESTIMATOR

2.1. Main idea

This work focuses on signals which can be decomposed into a sum of sinusoidal components and a colored noise. In the following, a moving average process is assumed for the latter, with a corresponding FIR filter of transfer function $B(z)$. The spectral envelope of the partials is modeled by an autoregressive filter of transfer function $\frac{1}{A(z)}$. The technique presented herein is based on the decomposition of the set of DFT frequencies into two subsets: the subset $\mathcal{N}$ owing to the background noise properties and the other, $\mathcal{H}$, associated with the sinusoidal part. Once both $1/A(z)$ and $B(z)$ have been estimated, the constructed likelihood is maximized for the true value of F_0 since it simultaneously whitens the noise sub-spectrum and the sinusoidal sub-spectrum. In the case where a bad F_0 candidate is selected, the choice of a FIR $\mathcal{N}$-support sub-spectrum and an AR $\mathcal{H}$-support sub-spectrum ensures that such a flatness of both sub-spectra is not achieved.

2.2. Statistical framework

Let $\boldsymbol{x}$ denote the N-dimensional vector containing N successive samples of data, $\boldsymbol{X}$ the N-dimensional vector of its Digital Fourier Transform (DFT) and $\boldsymbol{F}$ the $N \times N$ orthonormal DFT matrix ($F_{(p,q)} = \frac{1}{\sqrt{N}} e^{-2i\pi \frac{pq}{N}}$). We assume that $\boldsymbol{x}$ results from the circular filtering of a centered white complex Gaussian random vector $\boldsymbol{w}$ of variance σ^2. Let $\boldsymbol{h}$ be the corresponding impulse response vector, and $\boldsymbol{H}$ its N-dimensional DFT vector. Since $\boldsymbol{X} = \mathrm{diag}\{\boldsymbol{H}\}\boldsymbol{F}\boldsymbol{w}$, $\boldsymbol{X}$ is a centered Gaussian random vector of covariance matrix $\sigma^2 \,\mathrm{diag}\{|\boldsymbol{H}|^2\}$.

Below, we consider that the observed data consist of a subset $\mathcal{S}$ of the DFT coefficients in vector $\boldsymbol{X}$. Then the previous discussion shows that the probability law of the

observed data is

$$p(X_{\mathcal{S}}) = \prod_{k\in\mathcal{S}} \frac{1}{\pi\sigma^2|H(k)|^2} e^{-\frac{|X(k)|^2}{\sigma^2|H(k)|^2}}.$$

Thus the normalized log-likelihood $L_{\mathcal{S}}(\sigma, \boldsymbol{h}) = \frac{1}{\#\mathcal{S}} \ln p(X_{\mathcal{S}})$ can be written in the form

$$L_{\mathcal{S}}(\sigma, \boldsymbol{h}) = C + \frac{1}{\#\mathcal{S}} \sum_{k\in\mathcal{S}} \left[\ln\left(\frac{|X(k)|^2}{\sigma^2|H(k)|^2}\right) - \frac{|X(k)|^2}{\sigma^2|H(k)|^2} \right] \quad (1)$$

where $C = -\frac{1}{\#\mathcal{S}} \sum_{k\in\mathcal{S}} \ln(\pi|X(k)|^2)$ is a constant with respect to σ and $\boldsymbol{h}$, and $\#\mathcal{S}$ denotes the number of elements in $\mathcal{S}$. Normalizing the likelihood by factor $1/\#\mathcal{S}$ aims at obtaining comparable, homogeneous values when $\#\mathcal{S}$ varies. Maximizing $L_{\mathcal{S}}$ with respect to σ yields the estimate

$$\widehat{\sigma}^2 = \frac{1}{\#\mathcal{S}} \sum_{k\in\mathcal{S}} \left|\frac{X(k)}{H(k)}\right|^2. \quad (2)$$

Then substituting equation (2) into equation (1) yields

$$L_{\mathcal{S}}(\boldsymbol{h}) \triangleq L_{\mathcal{S}}(\widehat{\sigma}^2, \boldsymbol{h}) = C - 1 + \ln\left(\rho_{\mathcal{S}}(\boldsymbol{h})\right) \quad (3)$$

where

$$\rho_{\mathcal{S}}(\boldsymbol{h}) = \frac{\left(\prod_{k\in\mathcal{S}} \left|\frac{X(k)}{H(k)}\right|^2\right)^{\frac{1}{\#\mathcal{S}}}}{\frac{1}{\#\mathcal{S}} \sum_{k\in\mathcal{S}} \left|\frac{X(k)}{H(k)}\right|^2} \quad (4)$$

is equal to the ratio between the geometrical mean and the arithmetical mean of the set $\left\{\left|\frac{X(k)}{H(k)}\right|^2\right\}_{k\in\mathcal{S}}$. Such a ratio is maximal and equal to 1 when $|X(k)/H(k)|$ is constant, independant of k, which means that $\rho_{\mathcal{S}}(\boldsymbol{h})$ measures the *whiteness*, or the *flatness* of $\left\{\frac{X(k)}{H(k)}\right\}_{k\in\mathcal{S}}$. The next step consists in choosing a parametric model for $\boldsymbol{h}$, and maximizing $L_{\mathcal{S}}$ with respect to the filter parameters. This optimization results in maximizing $\rho_{\mathcal{S}}(\boldsymbol{h})$. For instance, if $\boldsymbol{h}$ is modeled as an autoregressive (AR) filter, an approximate solution $\widehat{\boldsymbol{h}}$ to the optimization problem can be obtained by means of linear prediction techniques [9]. If $\boldsymbol{h}$ is modeled as a finite impulse response (FIR) filter of length $p \ll N$, an approximate solution $\widehat{\boldsymbol{h}}$ can be obtained by windowing a biased estimate of the autocovariance function.

2.3. Application to pitch estimation

Our pitch estimator relies on a weighted maximum likelihood (WML) method: for all subsets $\mathcal{H}$, *i.e.* for all possible F_0's, we calculate the weighted likelihood

$$L_{\mathcal{H}} = \alpha \ln \hat{\rho_{\mathcal{H}}} + (1-\alpha) \ln \hat{\rho_{\mathcal{N}}} \quad (5)$$

$$\text{with } \begin{cases} \hat{\rho_{\mathcal{H}}} = \max\limits_{A} \rho_{\mathcal{H}}\left(\frac{1}{A(z)}\right) \\ \hat{\rho_{\mathcal{N}}} = \max\limits_{B} \rho_{\mathcal{N}}\left(B(z)\right) \end{cases}$$

where $\mathcal{N} = \overline{\mathcal{H}}$ is the complement set of $\mathcal{H}$ and $0 < \alpha < 1$ (in practice we choose $\alpha = 1/2$). The pitch estimate is given by the set $\hat{\mathcal{H}}$ which maximizes $L_{\mathcal{H}}$. This maximum depends on the sum of the two $\mathcal{H}$-dependent terms in (5): $\ln \hat{\rho_{\mathcal{H}}}$ and $\ln \hat{\rho_{\mathcal{N}}}$. The flatness $\hat{\rho_{\mathcal{H}}}$ of the whitened components has a local maximum for a smooth spectral envelope, obtained when analyzing the true F_0 (see figure 1) or one of its multiples (*i.e.* $\mathcal{H}$ is a subset of the right set of overtones, see figure 2), or when $\mathcal{H}$ only contains noisy components. Low values of $\hat{\rho_{\mathcal{H}}}$ are obtained when amplitudes at the frequencies of $\hat{\mathcal{H}}$ are alternately low and high since AR filters have no zero, which means that they cannot fit a spectrum where some sinusoidal components in $\mathcal{H}$ are missing. This particularly happens for a sub-harmonic of the true F_0 (see figure 3). In other respects, when considering the spectral envelope of the noisy part of the sound, FIR filters have no pole, which means that they cannot fit any sinusoidal component: the spectral flatness $\hat{\rho_{\mathcal{N}}}$ of the whitened residual part reaches high values when the frequencies of overtones have been selected in $\mathcal{H}$, *i.e.* when analyzing any sub-harmonic frequency of the true F_0 (see figure 3). As illustrated in figure 4, by combining both spectral flatnesses $\hat{\rho_{\mathcal{H}}}$ and $\hat{\rho_{\mathcal{N}}}$, a global maximum is found for the true F_0 while any other local maximum in $\hat{\rho_{\mathcal{H}}}$ (or $\hat{\rho_{\mathcal{N}}}$) is attenuated by $\hat{\rho_{\mathcal{N}}}$ (or $\hat{\rho_{\mathcal{H}}}$), particularly harmonics and sub-harmonics.

3. APPLICATION TO MULTI-PITCH ESTIMATION OF PIANO TONES

3.1. Inharmonicity in piano tones

In a piano note, the stiffness of strings causes the frequencies of overtones to slightly differ from a perfect harmonic distribution. We are focussing one these quasi-harmonic sounds and exclude from this study other inharmonic tones like bell tones. The frequency of the overtone of order n is thus given by the inharmonicity law [10]:

$$f_n^{(f_0,\beta)} = n f_0 \sqrt{1 + \beta\left(n^2 - 1\right)} \quad (6)$$

where f_0 is the fundamental frequency and β is the inharmonicity coefficient. Note that β varies along the range of the piano keyboard and from one instrument to the other. Thus, the set $\mathcal{H}$, characterized by these two parameters, is defined as:

$$\mathcal{H}^{(f_0,\beta)} = \left\{ f_n^{(f_0,\beta)} / n \in \mathbb{N}, f_n^{(f_0,\beta)} < F_s/2 \right\} \quad (7)$$

Analysis of a synthetic signal with fundamental frequency 1076.6602 Hz.

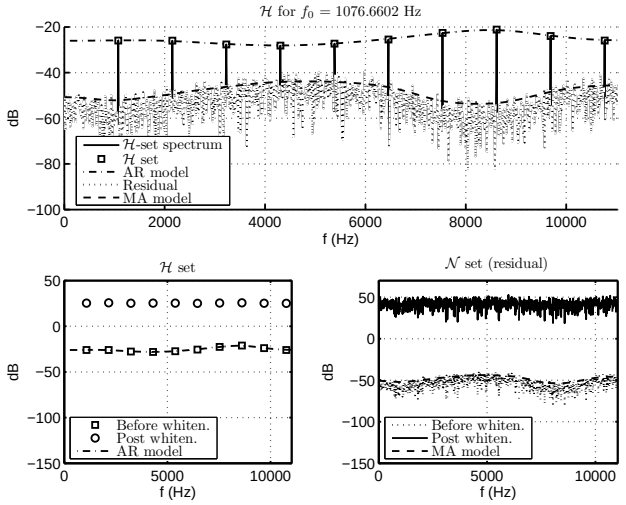

Figure 1: $L_{\mathcal{H}}$ estimation for $\mathcal{H} = \hat{\mathcal{H}}$ (true F_0). Overtones are selected in the spectrum (top), amplitudes of components fit the AR model (bottom left) and the residual spectrum is well whitened by the MA model (bottom right). In order to avoid overlapping between curves in the graphical representation, an constant offset is added to post-whitening dB-curves.

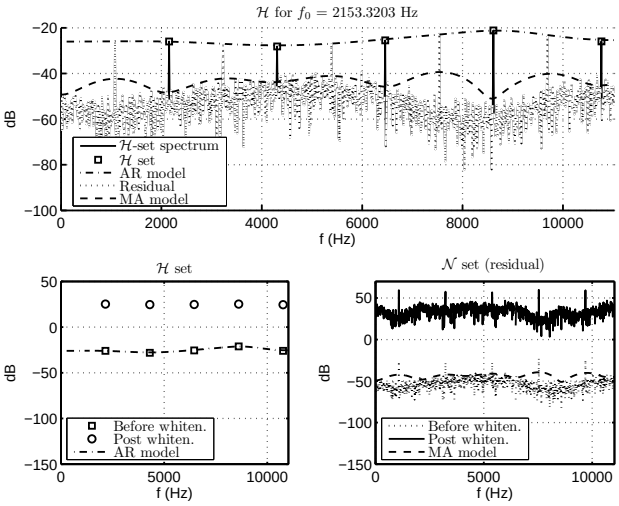

Figure 2: $L_{\mathcal{H}}$ estimation at twice the true F_0. Amplitudes of components fit the AR model whereas the residual spectrum is not perfectly whitened by the MA model, due to remaining components.

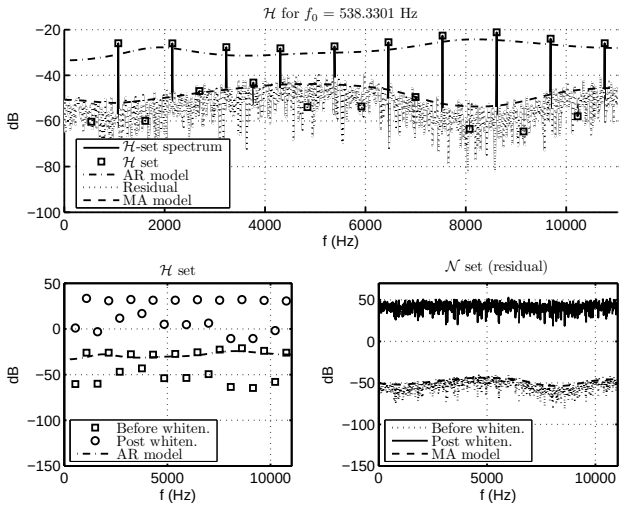

Figure 3: $L_{\mathcal{H}}$ estimation at half the true F_0. While residual spectrum is well whitened by the MA model, amplitudes of components do not fit the AR model, resulting in a low flatness of whitened amplitudes (bottom left, circles).

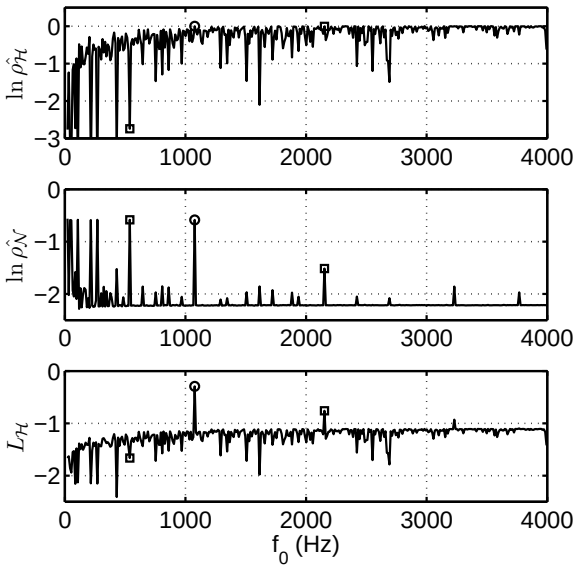

Figure 4: $\mathcal{H}$-dependent terms $\ln \hat{\rho_{\mathcal{H}}}$ (top) and $\ln \hat{\rho_{\mathcal{N}}}$ (middle), and weighted likelihood $L_{\mathcal{H}}$ (bottom), computed for all possible F_0's (*i.e.* all possible $\mathcal{H}$'s).

where F_s is the sampling frequency. Optimizing the log-likelihood $L\left(\mathcal{H}^{(f_0,\beta)}\right)$ with respect to $\mathcal{H}^{(f_0,\beta)}$ then consists in maximizing it with respect to f_0 and β.

3.2. From the theoretical model to real sounds

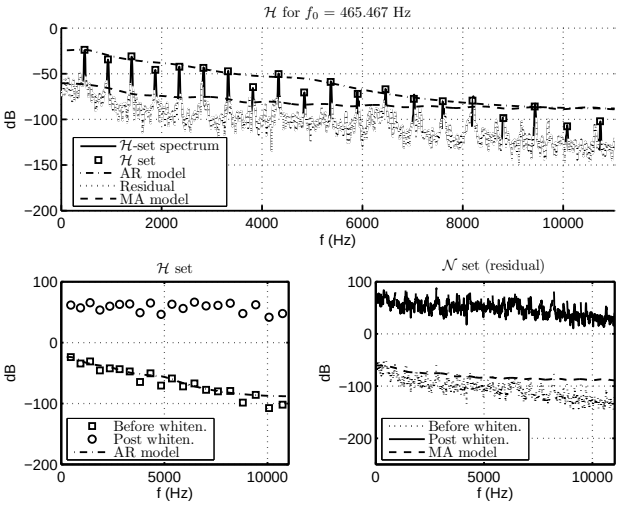

Figure 5: Real piano tone: separation between note components and residual part, and related MA and AR models

How do real piano tones fit the signal model described above? The AR model for the sinusoidal component, the MA noise model and the inharmonicity distribution of frequencies seem to be robust hypotheses. Conversely, the practical application of the method has to cope with two deviations from the theoretical point of view:

1. the assumption that f_n lies in the exact center of a frequency bin (multiple of $1/N$) is usually false, and spectral leakage thus influences the $\mathcal{N}$-support sub-spectrum.

2. the amplitude of the overtone may vary within the analysis frame, reflecting various effects as the energy loss of the sound and the beating between close adjacent components. This can affect the spectral envelope of the $\mathcal{H}$-support sub-spectrum.

The windowing of the analyzed waveform by a Hann window has proved to be a robust trade-off to overcome these issues. It prevents the spectral leakage associated with high energy components from masking weak overtones. Amplitudes of every overtone k are estimated by performing a parabolic interpolation of the spectrum (in decibels) based on the values in the nearest Fourier bins. The resulting (linear) value is used when computing the sinusoidal-part spectral flatness $\hat{\rho_{\mathcal{H}}}$, *i.e.* in place of $X(k)$ in equation (4). In order to minimize the effects described above in $\hat{\rho_{\mathcal{N}}}$ (see equation (4)), primary lobes of the frequencies selected in $\mathcal{H}$ are removed from $\mathcal{N}$, which is redefined as:

$$\mathcal{N} = \{k'/\forall f \in \mathcal{H}, |k'/N - f| > \Delta f/2\} \quad (8)$$

where Δf is the width of the primary lobe ($\Delta f = \frac{4}{N}$ for a Hann window). Note that the question of removing a set of components is a key step in the implementation of our algorithm. As shown in figure 5, the proposed method performs an approximate removal that offers a satisfying trade-off between efficiency and computational cost. Other techniques based on amplitude estimation and adapted filter design have been tested without bringing major improvements. The non-stationary nature of signals seems to be responsible for this limitation. It should be taken into account for enhancing the separation between a set of components and the residual signal.

3.3. Extension to polyphonic sounds

We now consider that the deterministic signal $s(n)$ is a sum of M inharmonic sounds: $s(n) = \sum_{m=1}^{M} s^{(m)}(n)$ and $\forall m \in \{1 \dots M\}$, $f_n^{(m)} = n f_0^{(m)} \sqrt{1 + \beta^{(m)}(n^2 - 1)}$, where $f_0^{(m)}$ is the pitch and $\beta^{(m)} > 0$ is the inharmonicity coefficient of the m^{th} tone. Each note is associated with one individual AR model, and weights in the likelihood are uniformly distributed among notes. Thus the WML principle consists in maximizing the log-likelihood:

$$L(\mathcal{H}^{(1)}, \dots, \mathcal{H}^{(M)}) = \frac{1}{2M} \sum_{m=1}^{M} \ln \rho_{\mathcal{H}^{(m)}} \left(\frac{1}{A^{(m)}(z)} \right) + \frac{1}{2} \ln \hat{\rho_{\mathcal{N}}} \quad (9)$$

where $\mathcal{H}^{(m)} = \mathcal{H}^{\left(f_0^{(m)}, \beta^{(m)}\right)}$ and $\mathcal{N}$ is the set of bins outside primary lobes of frequencies of any $\mathcal{H}^{(m)}$. The optimization is performed with respect to each of the sets $\mathcal{H}^{(1)}, \dots, \mathcal{H}^{(M)}$. Each set $\mathcal{H}^{(m)}$ is defined by the parameters $\{(f_0^{(m)}, \beta^{(m)})\}_{m \in \{1 \dots M\}}$ and $1/A^{(m)}(z)$ is the AR filter related to note m. Two distinct sets $\mathcal{H}^{(m_1)}$ and $\mathcal{H}^{(m_2)}$ may intersect, allowing overlap between spectra of notes m_1 and m_2. The algorithm presented in section 2.3 can be applied straightforwardly.

3.4. Multi-pitch estimator implementation

Multi-pitch estimation is often performed either in an iterative or in a joint process. The proposed method belongs to the joint estimation category. While iterative methods consist in successively estimating and removing a predominant F_0, joint estimation simultaneously extracts the set of

F_0's. Thus, a direct implementation of the algorithm described above would require to compute the ML of all possible combinations of notes, leading to a high-order combinatory task. For instance, more than 2.10^6 different chords exist for a 4-note polyphony in the full piano range, each of these candidates requiring several calls to the likelihood function since the exact F_0 and β values are unknown.

In order to reduce the cost of the ML estimation, a two-step algorithm is proposed. First, each possible chord is evaluated on a reduced number of points N_p in the $(f_0^{(m)}, \beta^{(m)})$ region around F_0 values from the well-tempered scale and approximate β values. N_{cand} chord candidates are extracted among all combinations by selecting the N_{cand} greatest likelihood values. Then, the likelihood of each selected candidate is locally maximized with respect to coefficients $f_0^{(m)}$ and $\beta^{(m)}$. A simplex method is used to perform this optimization, which is initialized with the $f_0^{(m)}$ and $\beta^{(m)}$ values selected during the first step. Finally, the chord with maximum accurately-computed likelihood is selected as the chord estimate.

4. EXPERIMENTAL RESULTS

The algorithm has been tested on a database composed of about 540 isolated piano tones of the RWC database [11] and random chords generated by several virtual piano softwares based on sampled sounds. About 600 two-note chords and 600 three-note chords were evaluated. In each case, the polyphony is known a priori by the algorithm and the estimation results from the analysis of one 93 ms frame, beginning 10 ms after the onset. F_0 estimates are rounded to the nearest half-tone in the well-tempered scale in order to determine if an estimated note is correct. This approximation on F_0 is carried out in order to evaluate the pitch estimation at a note level rather than at a frequency level. The note search range spreads over 5 octaves, from MIDI note 36 ($f_0 = 65$ Hz) to MIDI note 95 ($f_0 = 1976$ Hz). These test conditions are similar to the ones used in competitor systems [4, 5, 7] in terms of frame length, F_0 search range and error rate definition.

The parameters of the system have been adjusted as follows. Sounds are sampled at 22050 Hz. DFT are computed on 4096 points after zero-padding the 2048-point frame. The AR model order is set to 8, the MA model order to 20. In the first step of the implementation described in section 3.4, all chord combinations are evaluated, each one with $N_p = 10$ (polyphony ≤ 2) or $N_p = 5$ (polyphony three) different $(f_0^{(m)}, \beta^{(m)})$ values. Then $N_{cand} = 75$ (monophony) or $N_{cand} = 150$ (polyphony ≥ 2) chord candidates are selected for the second step.

Error rates are 2.0% in monophony, 7.5% in polyphony two and 23.9% in polyphony three. They are reported in

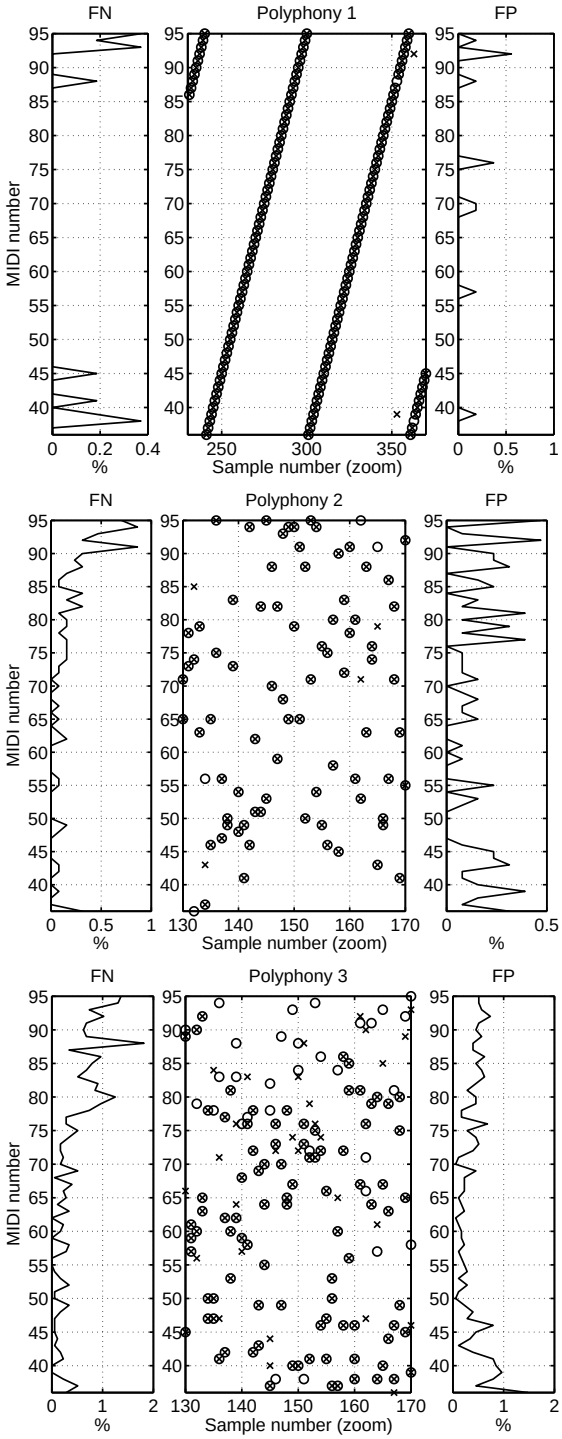

Figure 6: Estimation results: for a given polyphony (1 to 3 from top to bottom), random chords are generated (circles) and estimated (crosses). For visual representation clarity, only 50 samples of them are shown (center). Distribution of false negatives is displayed on the left. Distribution of false positives is displayed on the right.

Polyphony	1	2	3
Error rate	2.0% $\pm 0.6\%$	7.5% $\pm 0.75\%$	23.9% $\pm 1.0\%$
Octave error rate	0%	1.6%	5.2%
State of the art	$2 \sim 11\%$	$7 \sim 25\%$	$\approx 10 \sim 35\%$

Table 1: Error rates with respect to polyphony. Lower and upper bounds of state-of-the-art performances are also reported. Confidence interval is derived as the standard deviation of the error rate estimator.

table 1 and can be compared to the three competitor systems previously mentionned. Their performances have been established in [7] for polyphony one, two, four and six: error rates vary from 2 to 11% in monophony, from 7 to 25% in polyphony two and from 14 to 41% in polyphony four. Error rates in polyphony three are not given, but could be figured out as intermediate values between results in polyphony two and four, which would lead to approximate error rates between 10 and 35%. The proposed pitch estimator is comparable to competitor systems in terms of performance. Error rates are particularly competitive in polyphonies one and two.

The evaluation task has been performed using randomly uniformly-distributed notes in order to provide experimental results from an objective point of view rather than from musical considerations. The distribution of errors is reported in figure 6. The few errors in polyphony one occur in the lowest and highest pitch regions. In polyphony two and three, most of missed notes (or false negatives, FN) are located in the treble part of the piano range whereas the false-alarm notes (or false positives, FP) estimated in place of them tend to be distributed in a more uniform manner along the piano range. Closely-spaced chords in the medium range seem easier to detect than widely-spaced chords. Octave error are scarce – around one fifth of all errors for each polyphony number –, which can be explained by the complementary contributions of note and noise likelihoods. On the contrary, high-pitched FN and large-interval errors often occur, in spite of the likelihood normalization stage, due to the sensitivity of the ML approach to the variable number of frequency parameters that depends on F_0 candidates.

5. CONCLUSIONS

The multipitch estimation task has been performed here through a Maximum Likelihood approach. It consists in modeling notes and residual noise by AR and MA models, and results in a criterion on their spectral flatness after a whitening process based on the models. The method has been validated by satisfying experimental results for polyphony one to three.

Future works will deal with managing the overlap between notes spectra, with improving the model for the spectral envelope of notes and with making the computational cost decrease in order to both benefit from the efficiency of the estimator and avoid the inherent complexity of joint estimation of multiple F_0's.

6. REFERENCES

[1] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, 2002.

[2] M. Ryynänen and A.P. Klapuri, "Polyphonic music transcription using note event modeling," in *Proc. of WASPAA*, New Paltz, NY, USA, October 2005, IEEE, pp. 319–322.

[3] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. on Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

[4] 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.

[5] A.P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, November 2003.

[6] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. of ICASSP 2006*, Toulouse, France, May 14-29 2006, IEEE, vol. 5, pp. 53–56.

[7] A.P. Klapuri, "A perceptually motivated multiple-f0 estimation method," in *Proc. of WASPAA*, New Paltz, NY, USA, October 2005, IEEE, pp. 291– 294.

[8] Manuel Davy, Simon Godsill, and Jerome Idier, "Bayesian analysis of polyphonic western tonal music," *JASA*, vol. 119, no. 4, pp. 2498–2517, 2006.

[9] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[10] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer, 1998.

[11] T. Nishimura M. Goto, H. Hashiguchi and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. of ISMIR*, Baltimore, Maryland, USA, 2003, pp. 229–230.