A PROBABILISTIC LINE SPECTRUM MODEL FOR MUSICAL INSTRUMENT SOUNDS AND ITS APPLICATION TO PIANO TUNING ESTIMATION.

François Rigaud[1*], *Angélique Drémeau*[1], *Bertrand David*[1] *and Laurent Daudet*[2†]

[1] Institut Mines-Télécom; Télécom ParisTech; CNRS LTCI; Paris, France
[2] Institut Langevin; Paris Diderot Univ.; ESPCI ParisTech; CNRS; Paris, France

ABSTRACT

The paper introduces a probabilistic model for the analysis of line spectra – defined here as a set of frequencies of spectral peaks with significant energy. This model is detailed in a general polyphonic audio framework and assumes that, for a time-frame of signal, the observations have been generated by a mixture of notes composed by partial and noise components. Observations corresponding to partial frequencies can provide some information on the musical instrument that generated them. In the case of piano music, the fundamental frequency and the inharmonicity coefficient are introduced as parameters for each note, and can be estimated from the line spectra parameters by means of an Expectation-Maximization algorithm. This technique is finally applied for the unsupervised estimation of the tuning and inharmonicity along the whole compass of a piano, from the recording of a musical piece.

Index Terms— probabilistic model, EM algorithm, polyphonic piano music

1. INTRODUCTION

Most algorithms dedicated to audio applications (F_0-estimation, transcription, ...) consider the whole range of audible frequencies to perform their analysis, while besides attack transients, the energy of music signals is often contained into only a few frequency components, also called partials. Thus, in a time-frame of music signal only a few frequency-bins carry information relevant for the analysis. By reducing the set of observations, *ie.* by keeping only the few most significant frequency components, it can be assumed that most signal analysis tasks may still be performed. For a given frame of signal, this reduced set of observations is here called a line spectrum, this appellation being usually defined for the discrete spectrum of electromagnetic radiations of a chemical element.

Several studies have considered dealing with these line spectra to perform analysis. Among them, [1] proposes to compute tonal descriptors from the frequencies of local maxima extracted from polyphonic audio short-time spectra. In [2] a probabilistic model for multiple-F_0 estimation from sets of maxima of the Short-Time Fourier Transform is introduced. It is based on a Gaussian mixture model having means constrained by a F_0 parameter and solved as a maximum likelihood problem by means of heuristics and grid search. A similar constrained mixture model is proposed in [3] to model speech spectra (along the whole frequency range) and solved using an Expectation-Maximization (EM) algorithm.

*This work is supported by the DReaM project of the French Agence Nationale de la Recherche (ANR-09-CORD-006, under CONTINT program).

†Also at Institut Universitaire de France.

The model presented in this paper is inspired by these two last references [2, 3]. The key difference is that we here focus on piano tones, which have the well-known property of inharmonicity, that in turn influences tuning. This slight frequency stretching of partials should allow, up to a certain point, disambiguation of harmonically-related notes. Reversely, from the set of partials frequencies, it should be possible to estimate the inharmonicity and tuning parameters of the piano. The model is first introduced in a general audio framework by considering that the frequencies corresponding to local maxima of a spectrum have been generated by a mixture of notes, each note being composed of partials (Gaussian mixture) and noise components. In order to be applied to piano music analysis, the F_0 and inharmonicity coefficient of the notes are introduced as constraints on the means of the Gaussians and a maximum *a posteriori* EM algorithm is derived to perform the estimation. It is finally applied to the unsupervised estimation of the inharmonicity and tuning curves along the whole compass of a piano, from isolated note recordings, and then from a polyphonic piece.

2. MODEL AND PROBLEM FORMULATION

2.1. Observations

In time-frequency representations of music signals, the information contained in two consecutive frames is often highly redundant. This suggests that in order to retrieve the tuning of a given instrument from a whole piece of solo music, a few independent frames localized after note onset instants should contain all the information that is necessary for processing. These time-frames are indexed by $t \in \{1...T\}$ in the following. In order to extract significant peaks (*i.e.* peaks containing energy) from the magnitude spectra a noise level estimation based on median filtering (*cf.* appendix of [4]) is first performed. Above this noise level, local maxima (defined as having a greater magnitude than K left and right frequency bins) are extracted. The frequency of each maximum picked in a frame t is denoted by y_{ti}, $i \in \{1...I_t\}$. The set of observations for each frame is then denoted by $\mathbf{y}_t$ (a vector of length I_t), and for the whole piece of music by $Y = \{\mathbf{y}_t,\ t \in \{1...T\}\}$. In the following of this document, the variables denoted by lower case, bold lower case and upper case letters will respectively correspond to scalars, vectors and sets of vectors.

2.2. Probabilistic Model

If a note of music, indexed by $r \in \{1...R\}$, is present in a time-frame, most of the extracted local maxima should correspond to partials related by a particular structure (harmonic or inharmonic for instance). These partial frequencies correspond to the set of

parameters of the proposed model. It is denoted by θ, and in a general context (no information about the harmonicity or inharmonicity of the sounds) can be expressed by $\theta = \{f_{nr} | \forall n \in \{1...N_r\}, r \in \{1...R\}\}$, where n is the rank of the partial and N_r the maximal rank considered for the note r.

In order to link the observations to the set of parameter θ, the following hidden random variables are introduced:
• $q_t \in \{1...R\}$, corresponding to the note that could have generated the observations $\mathbf{y}_t$.
• $C_t = [c_{tir}]_{(i,r)\in\{1...I_t\}\times\{1...R\}}$ gathering Bernoulli variables specifying the nature of the observation y_{ti}, for each note r. An observation is considered belonging to the partial of a note r if $c_{tir} = 1$, or to noise (non-sinusoidal component or partial corresponding to another note) if $c_{tir} = 0$.
• $P_t = [p_{tir}]_{(i,r)\in\{1...I_t\}\times\{1...R\}}$ corresponding to the rank of the partial n of the note r that could have generated the observation y_{ti} provided that $c_{tir}=1$.

Based on these definitions, the probability that an observation y_{ti} has been generated by a note r can be expressed as:

$$\begin{aligned} p(y_{ti}|q_t=r;\theta) &= p(y_{ti}|c_{tir}=0, q_t=r) \cdot p(c_{tir}=0|q_t=r) \\ &+ \sum_n p(y_{ti}|p_{tir}=n, c_{tir}=1, q_t=r;\theta) \\ &\cdot \; p(p_{tir}=n|c_{tir}=1, q_t=r) \cdot p(c_{tir}=1|q_t=r). \end{aligned} \quad (1)$$

It is chosen that the observations that are related to the partial n of a note r should be located around the frequencies f_{nr} according to a Gaussian distribution of mean f_{nr} and variance σ_r^2 (fixed parameter):

$$p(y_{ti}|p_{tir}=n, c_{tir}=1, q_t=r;\theta) = \mathcal{N}(f_{nr}, \sigma_r^2), \quad (2)$$

$$p(p_{tir}=n|c_{tir}=1, q_t=r) = 1/N_r. \quad (3)$$

On the other hand, observations that are related to noise are chosen to be uniformly distributed along the frequency axis (with maximal frequency F):

$$p(y_{ti}|c_{tir}=0, q_t=r) = 1/F. \quad (4)$$

Then, the probability to obtain a noise or partial observation knowing the note r is chosen so that:
· if $I_t > N_r$:

$$p(c_{tir}|q_t=r) \underset{I_t>N_r}{=} \begin{cases} (I_t - N_r)/I_t & \text{if } c_{tir} = 0, \\ N_r/I_t & \text{if } c_{tir} = 1. \end{cases} \quad (5)$$

This should approximately correspond to the proportion of observations associated to noise and partial classes for each note.
· if $I_t \leq N_r$:

$$p(c_{tir}|q_t=r) \underset{I_t\leq N_r}{=} \begin{cases} 1-\epsilon & \text{if } c_{tir} = 0, \\ \epsilon & \text{if } c_{tir} = 1, \end{cases} \quad (6)$$

with $\epsilon \ll 1$ (set to 10^{-5} in the presented results). This latter expression means that for a given note r at a frame t, every observation should be mainly considered as noise if N_r (its number of partials), is greater than the number of observations. This situation may occur for instance in a frame in which a single note from the high treble range is played. In this case, only a few local maxima are extracted and lowest notes, composed of much more partials, should not be considered as present.

Finally, with no prior information it is chosen

$$p(q_t=r) = 1/R. \quad (7)$$

2.3. Estimation problem

In order to estimate the parameters of interest θ, it is proposed to solve the following maximum *a posteriori* estimation problem:

$$(\theta^\star, \{C_t^\star\}_t, \{P_t^\star\}_t) = \underset{\theta, \{C_t\}_t, \{P_t\}_t}{\operatorname{argmax}} \sum_t \log p(\mathbf{y}_t, C_t, P_t; \theta), \quad (8)$$

where

$$p(\mathbf{y}_t, C_t, P_t; \theta) = \sum_r p(\mathbf{y}_t, C_t, P_t, q_t = r; \theta). \quad (9)$$

Solving problem (8) corresponds to the estimation of θ, joint to a clustering of each observation into noise or partial classes for each note. Note that the sum over t of Eq. (8) arises from the time-frame independence assumption (justified in Sec. 2.1).

2.4. Application to piano music

The model presented in Sec. 2.2 is general since no particular structure has been set on the partial frequencies. In the case of piano music, the tones are inharmonic and the partials frequencies related to transverse vibrations of the (stiff) strings can be modeled as:

$$f_{nr} = nF_{0r}\sqrt{1 + B_r n^2}, \;\; n \in \{1...N_r\}. \quad (10)$$

F_{0r} corresponds to the fundamental frequency (theoretical value, that does not appear as one peak in the spectrum) and B_r to the inharmonicity coefficient. These parameters vary along the compass and are dependent on the piano type [5]. Thus, for applications to piano music, the set of parameters can be rewritten as $\theta = \{F_{0r}, B_r, \forall r \in \{1, R\}\}$.

3. OPTIMIZATION

Problem (8) has usually no closed-form solution but can be solved in an iterative way by means of an Expectation-Maximization (EM) algorithm [6]. The auxiliary function at iteration $(k+1)$ is given by

$$\begin{aligned} Q(\theta, \{C_t\}_t, \{P_t\}_t | \theta^{(k)}, \{C_t^{(k)}\}_t, \{P_t^{(k)}\}_t) = \\ \sum_t \sum_r \omega_{rt} \cdot \sum_i \log p(y_{ti}, c_{tir}, p_{tir}, q_t=r; \theta) \end{aligned} \quad (11)$$

where,

$$\omega_{rt} \triangleq p(q_t=r|\mathbf{y}_t, \{C_t^{(k)}\}_t, \{P_t^{(k)}\}_t; \theta^{(k)}), \quad (12)$$

is computed at the E-step knowing the values of the parameters at iteration (k). At the M-step, θ, $\{C_t\}_t$, $\{P_t\}_t$ are estimated by maximizing Eq. (11). Note that the sum over i in Eq. (11) is obtained under the assumption that in each frame the y_{ti} are independent.

3.1. Expectation

According to Eq. (12) and model Eq. (1)-(7)

$$\begin{aligned} \omega_{rt} &\propto \prod_{i=1}^{I_t} p(y_{ti}, q_t=r, c_{tri}^{(k)}, p_{tri}^{(k)}; \theta^{(k)}) \\ &\propto p(q_t=r) \cdot \prod_{i/\; c_{tir}^{(k)}=0} p(y_{ti}|q_t=r, c_{tir}^{(k)}) \cdot p(c_{tir}^{(k)}|q_t=r) \\ &\cdot \prod_{i/\; c_{tir}^{(k)}=1} p(y_{ti}|q_t=r, c_{tir}^{(k)}, p_{tir}^{(k)}, \theta^{(k)}) \cdot p(p_{tir}^{(k)}|c_{tir}^{(k)}, q_t=r) \cdot p(c_{tir}^{(k)}|q_t=r), \end{aligned} \quad (13)$$

normalized so that $\sum_{r=1}^R \omega_{rt} = 1$ for each frame t.

3.2. Maximization

The M-step is performed by a sequential maximization of Eq. (11):

• First, estimate $\forall\ t, i$ and $q_t = r$ the variables c_{tir} and p_{tir}. As mentioned in Sec. 2.3, this corresponds to a classification step, where each observation is associated, for each note, to noise class ($c_{tir} = 0$) or partial class with a given rank ($c_{tir} = 1$ and $p_{itr} \in \{1...N_r\}$). This step is equivalent to a maximization of $\log p(y_{ti}, c_{tir}, p_{tir} \,|\, q_t = r; \theta)$ which, according to Eq. (1)-(7), can be expressed as:

$$(c_{tir}^{(k+1)}, p_{tir}^{(k+1)}) = \tag{14}$$
$$\underset{(\{0,1\},n)}{\operatorname{argmax}} \begin{cases} -\log F + \log p(c_{tir}{=}0 | q_t{=}r), \\ \frac{-(y_{it} - f_{nr})^{.2}}{2\sigma_r^2} - \log N_r \sqrt{2\pi}\sigma_r + \log p(c_{tir}{=}1 | q_t{=}r). \end{cases}$$

• Then, the estimation of θ is equivalent to ($\forall r \in \{1...R\}$)

$$(F_{0r}^{(k+1)}, B_r^{(k+1)}) = \underset{F_{0r}, B_r}{\operatorname{argmax}} \sum_t \omega_{rt} \sum_{i/c_{tir}^{(k+1)}=1} \Big[\log p(c_{tir}^{(k+1)}{=}1 | q_t{=}r) - \left(y_{ti} - p_{tir}^{(k+1)} F_{0r} \sqrt{1 + B_r\, {p_{tir}^{(k+1)}}^2} \right)^2 \Big] \tag{15}$$

For F_{0r}, canceling the partial derivative of Eq. (15) leads to the following update rule:

$$F_{0r}^{(k+1)} = \frac{\sum_t \omega_{rt} \sum_{i/c_{tir}^{(k+1)}=1} y_{ti} \cdot p_{tir}^{(k+1)} \cdot \sqrt{1 + B_r\, {p_{tir}^{(k+1)}}^2}}{\sum_t \omega_{rt} \sum_{i/c_{tir}^{(k+1)}=1} {p_{tir}^{(k+1)}}^2 \cdot (1 + B_r\, {p_{tir}^{(k+1)}}^2)}. \tag{16}$$

For B_r, no closed-form solution can be obtained from the partial derivative of Eq. (15). The maximization is thus performed by means of an algorithm based on the Nelder-Mead simplex method.

3.3. Practical considerations

The cost-function (*cf.* maximization Eq. (8)) is non-convex with respect to (B_r, F_{0r}) parameters. In order to prevent the algorithm from converging towards a local maximum, a special care must be taken to the initialization.

First, the initialization of (B_r, F_{0r}) uses a mean model of inharmonicity and tuning [5] based on piano string design and tuning rule invariants. This initialization can be seen, depicted as gray lines, on Fig. 1(b) and 1(c) of Sec. 4. Moreover, to avoid situations where the algorithm optimizes the parameters of a note in order to fit the data corresponding to another note (*eg.* increasing F_0 of one semi-tone), (B_r, F_{0r}) are prevented from being updated over limit curves. For B, these are depicted as gray dashed-line in Fig. 1(b). The limits curves for F_0 are set to $+/-$ 40 cents of the initialization.

Since the deviation of the partial frequencies is increasing with the rank of partial (*cf.* Eq. (10)), the highest the rank of the partial, the less precise its initialization. Then, it is proposed to initialize the algorithm with a few partials for each note (about 10 in the bass range to 3 in the treble range) and to add a new partial every 10 iterations (number determined empirically) by initializing its frequency with the current (B_r, F_{0r}) estimates.

4. APPLICATIONS TO PIANO TUNING ESTIMATION

It is proposed in this section to apply the algorithm to the estimation of (B_r, F_{0r}) parameters from isolated note recordings covering the whole compass of pianos and polyphonic pieces, in an unsupervised way (*i.e.* without knowing which notes are played). The recordings are taken from *SptkBGCl* grand piano synthesizer (using high quality samples) of MAPS database[1].

The observation set is built according to the description given in Sec. 2.1. The time-frames are extracted after note onset instants and their length is set to 500 ms in order to have a sufficient spectral resolution. The FFT is computed on 2^{15} bins and the maxima are extracted by setting $K = 20$. Note that for the presented results, the knowledge of the note onset instants is taken from the ground truth (MIDI aligned files). For a complete blind approach, an onset detection algorithm should be first run. This should not significantly affect the results that are presented since onset detection algorithms usually perform well on percussive tones. Parameter σ_r is set to 2 Hz for all the notes and N_r maximal value is set to 40.

4.1. Estimation from isolated notes

The ability of the model/algorithm to provide correct estimates of (B_r, F_{0r}) on the whole piano compass is investigated here. The set of observations is composed of 88 frames (jointly processed), one for each note of the piano (from A0 to C8, with MIDI index in $[21, 108]$). R is set equal to 88 in order to consider all notes. The results are presented on Fig. 1. Subplot (a) depicts the matrix ω_{rt} in linear and decimal log. scale (x and y axis respectively correspond to the frame index t and note r in MIDI index). The diagonal structure can be observed up to frame $t{=}65$: the algorithm detected the good note in each frame, up to note $C\sharp 6$ (MIDI index 85). Above, the detection is not correct and leads to bad estimates of B_r (subplot (b)) and F_{0r} (subplot (c)). For instance, above MIDI note 97, (B_r, F_{0r}) parameters stayed fixed to their initial values. These difficulties in detecting and estimating the parameters for these notes in the high treble range are common for piano analysis algorithms [5]: in this range, notes are composed of 3 coupled strings that produce partials that do not fit well into the inharmonicity model Eq. (10). The consistency of the presented results may be qualitatively evaluated by refering to the curves of (B, F_0) obtained on the same piano by a supervised method, as depicted in Fig. 5 from [5].

4.2. Estimation from musical pieces

Finally, the algorithm is applied to an excerpt of polyphonic music (25 s of *MAPS_MUS-muss_3_SptkBGCl* file) containing notes in the range $D\sharp 1$- $F\sharp 6$ (MIDI 27-90) from which 46 frames are extracted. 66 notes, from A0 to C7 (MIDI 21-96), are considered in the model. This corresponds to a reduction of one octave in the high treble range where the notes, rarely used in a musical context, cannot be properly processed, as seen in Sec. 4.1.

The proposed application is here the learning of the inharmonicity and tuning curves along the whole compass of a piano from a generic polyphonic piano recording. Since the 88 notes are never present in a single recording, we estimate (B, F_0) for the notes present in the recording and, from the most reliable estimates, apply an interpolation based on physics/tuning considerations [5]. In order to perform this complex task in an unsupervised way, an heuristic is added to the optimization and a post-processing is performed. At each iteration of the optimization, a threshold is applied to ω_{rt} in order to limit the degree of polyphony to 10 notes for each frame t. Once the optimization is performed, the most reliable notes are kept according to two criteria. First, a threshold is applied to the matrix ω_{rt} so that elements having values lower than 10^{-3} are set

[1] http://www.tsi.telecom-paristech.fr/aao/en/category/database/

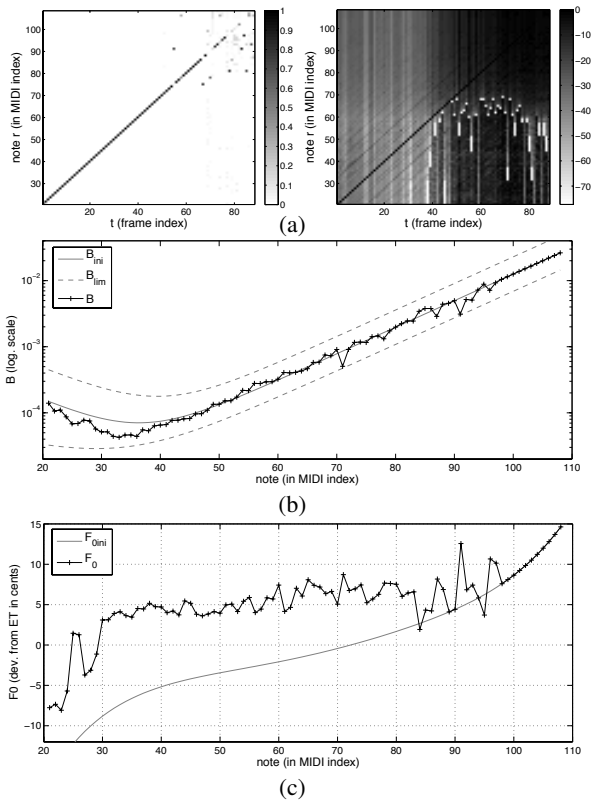

Figure 1: Analysis on the whole compass from isolated note recordings. a) ω_{rt} in linear (left) and $\log_{10}$ (right) scale. b) B in log scale and c) F_0 as deviation from Equal Temperament (ET) in cents, along the whole compass. (B, F_0) estimates are depicted as black '+' markers and their initialization as gray lines. The limits for the estimation of B are depicted as gray dashed-lines.

to zero. Then, notes that are never activated along the whole set of frames are rejected. Second, notes having B estimates stuck to the limits (*cf.* gray dashed lines in Fig. 1) are rejected.

Subplot 2(a) depicts the result of the note selection (notes having been detected at least once) for the considered piece of music. A frame-wise evaluation (with MIDI aligned) returned a precision of 86.4 % and a recall of 11.6 %, all notes detected up to MIDI index 73 corresponding to true positives, and above to false positives, all occuring in a single frame. It can be seen in subplots (b) and (c) that most of (B, F_0) estimates ('+' markers) corresponding to notes actually presents are consistent with those obtained from the single note estimation (gray lines). Above MIDI index 73, detected notes correspond to false positive and logically lead to bad estimates of (B, F_0). Finally, the piano tuning model [5] is applied to interpolate (B, F_0) curves along the whole compass (black dashed lines, indexed by WC) giving a qualitative agreement with the reference measurements. Note that bad estimates of notes above MIDI index 73 do not disturb the whole compass model estimation. Further work will address the quantitative evaluation of (B, F_0) estimation from synthetic signals, and real piano recordings (from which the reference has to be extracted manually [5]).

5. CONCLUSION

A probabilistic line spectrum model and its optimization algorithm have been presented in this paper. To the best of our knowledge,

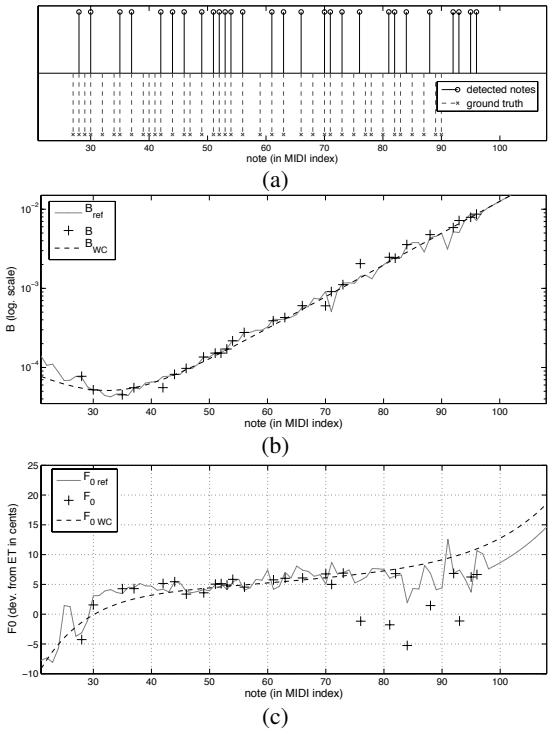

Figure 2: Piano tuning estimation along the whole compass from a piece of music. a) Note detected by the algorithm and ground truth. b) B in log scale. c) F_0 as deviation from ET in cents. (B, F_0) estimates are depicted as black '+' markers and compared to isolated note estimates (gray lines, obtained in Fig. 1). The interpolated curves (indexed by WC) are depicted as black dashed lines.

this is the only unsupervised estimation of piano inharmonicity and tuning estimation on the whole compass, from a generic extract of polyphonic piano music. Interestingly, for this task a perfect transcription of the music does not seem necessary: only a few reliable notes may be sufficient. However, an extension of this model to piano transcription could form a natural extension, but would require a more complex model taking account both temporal dependencies between frames, and spectral envelopes.

6. REFERENCES

[1] E. Gómez, "Tonal description of polyphonic audio for music content processing," *INFORMS Journal on Computing, Special Cluster on Computation in Music*, vol. 18, pp. 294–304, 2006.

[2] B. Doval and X. Rodet, "Estimation of fundamental frequency of musical sound signals," in *Proc. ICASSP*, April 1991.

[3] H. Kameoka, T. Nishimoto, and S. Sagayama, "Multi-pitch detection algorithm using constrained gaussian mixture model and information criterion for simultaneous speech," in *Proc. Speech Prosody (SP2004)*, March 2004, pp. 533–536.

[4] F. Rigaud, B. David, and L. Daudet, "A parametric model of piano tuning," in *Proc. DAFx'11*, September 2011.

[5] ——, "A parametric model and estimation techniques for the inharmonicity and tuning of the piano," *J. of the Acoustical Society of America*, vol. 133, no. 5, pp. 3107–3118, May 2013.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society (B),*, vol. 39, no. 1, 1977.