

Poisson point process modeling for polyphonic music transcription

Paul Peeling, Chung-fai Li, and Simon Godsill

Signal Processing Group, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, United Kingdom
php23@eng.cam.ac.uk, cfli@cuhk.edu.hk, sjg@eng.cam.ac.uk

Abstract: Peaks detected in the frequency domain spectrum of a musical chord are modeled as realizations of a nonhomogeneous Poisson point process. When several notes are superimposed to make a chord, the processes for individual notes combine to give another Poisson process, whose likelihood is easily computable. This avoids a data association step linking individual harmonics explicitly with detected peaks in the spectrum. The likelihood function is ideal for Bayesian inference about the unknown note frequencies in a chord. Here, maximum likelihood estimation of fundamental frequencies shows very promising performance on real polyphonic piano music recordings.

© 2007 Acoustical Society of America

PACS numbers: 43.60.Uv, 43.75.Xz, 43.60.Pt [JC]

Date Received: January 10, 2007 **Date Accepted:** February 4, 2007

1. Introduction

Music transcription refers to the generation of a musical score from audio data. The transcription of polyphonic music is of particular interest because of the underlying quasiperiodic structure of the individual note components. To exploit this, models describing the fundamental frequency of a note and its harmonics have been proposed.¹⁻³ Only a relatively small number of parameters is needed for a plausible description of the frequency content of the music. Often the performance of these schemes has been limited because the harmonics of different notes coincide, thus obscuring some of the components present in the data. Methods which account for this tend to either increase the model complexity² or use a heuristic approach to recover the missing components.⁴

Here we focus on determining multiple pitches within short frames of polyphonic music. As in many such systems,⁴⁻⁷ a preprocessing stage is assumed which extracts, or “detects,” individual peaks from a short-time frequency representation of the music. These peaks in the time-frequency domain are then modeled in a novel way as a nonhomogeneous Poisson point process.⁸ In such a model, the number of peaks detected in each frame is a Poisson random variable. In this formulation a likelihood function may be directly formulated for the observed data, without resorting to any data association task which assigns individual detected peaks to particular note fundamentals or harmonics.^{5,6} Thus we avoid the computational complexity of a full probabilistic data association, and also the heuristic approximations of suboptimal data association schemes.

The paper is organized as follows. In Sec. II we describe the basic Poisson process model for musical note clusters. Section III describes the estimation of rate functions for the model. Section IV describes a basic algorithm for implementation of the approach and Sec. V gives some results of application to real musical extracts. Finally, Sec. VI gives concluding remarks and points toward future developments.

2. Poisson point processes

Suppose that M simultaneous pitches are present in a frame of audio, with fundamental frequencies $F = \{f_1, f_2, \dots, f_M\}$, and this frame exhibits a number of peaks $P = \{p_1, p_2, \dots\}$ in the frequency domain. The k th element p_k in P is the frequency in Hz of a single peak in the spec-

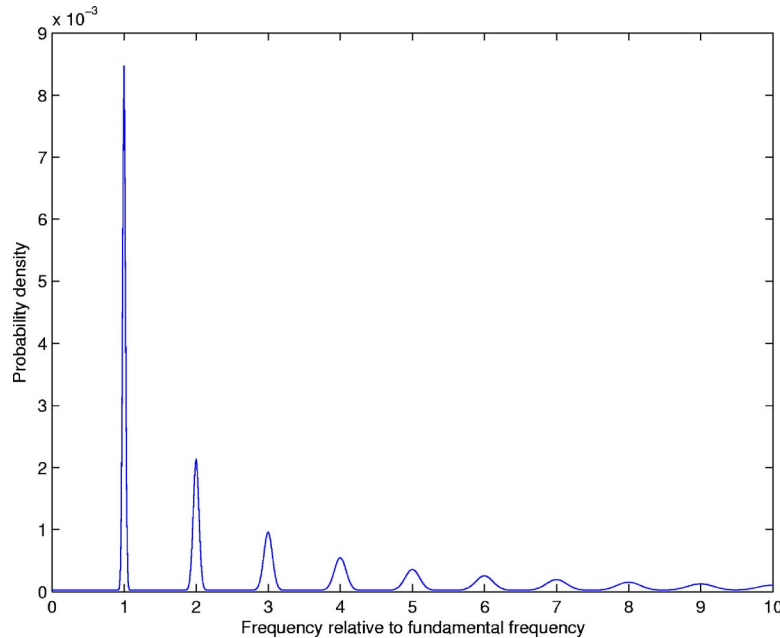


Fig. 1. (Color online) Detection of peaks from raw spectral amplitude data and representation as a point process on the frequency axis.

trum. It is expected that each note f_m will contribute peaks corresponding to its own harmonic structure. Any noise in the signal or inadequacies in the peak detection method may generate additional noise or “clutter” peaks (see Fig. 1).

Our fundamental assumption is that the peaks generated in this fashion are realizations from a *nonhomogeneous Poisson point process*⁸ with intensity function $\rho(p|F)$. The number of detected peaks $N_{\mathcal{A}}$ in an interval \mathcal{A} of the frequency domain is a Poisson random variable

$$p(N_{\mathcal{A}} = n) = \frac{e^{-\mu_{\mathcal{A}}}\mu_{\mathcal{A}}^n}{n!},$$

where the expected number of peaks in interval \mathcal{A} is

$$\mu_{\mathcal{A}} = \int_{\mathcal{A}} \rho(p|F) dp.$$

All peaks generated by the pitches have been combined into one set—we have technically taken a union of the individual point processes for each pitch. The union process of a number of independent Poisson processes is also a Poisson process, with intensity function given by the sum of the individual intensity functions. Therefore, we can decompose the intensity function into individual note and clutter components, i.e.

$$\rho(p|F) = \rho(p|f_1, f_2, \dots, f_M) = \rho_C(p) + \sum_{m=1}^M \rho(p|f_m), \tag{1}$$

where $\rho_C(p)$ is the predefined intensity function of the clutter process, and $\rho(p|f_m)$ is the intensity function of an individual note f_m .

It should be noted at this stage that this type of model is subtly different from those usually employed in peak modeling, which assume that each harmonic of each note in the spectrum has to be uniquely associated with at most one detected peak,^{6,9} which can lead to a com-

binomial explosion of data association terms when many notes are superimposed. Here, however, each harmonic may generate any number of detected peaks, in accordance with the intensity function $\rho(p|f_m)$. This will lead to substantial simplifications in computation. It also models the fact that individual harmonics may often lead to “split” peaks where several peaks are detected rather than just one. In a tracking setting a similar principle has recently been applied for simplification of the classical data association problem.^{10–12}

Peak detections will usually be made over a discrete set of frequency bins. Let $N(k)$ be the number of peaks occurring in the frequency interval $(k\Delta, (k+1)\Delta]$, where Δ is the frequency analysis bin size (easily made variable with frequency in multiscale approaches if required). Then, under the nonhomogeneous Poisson point process assumption, the probability of $N(k)$ peaks occurring is given by

$$P(N(k) = n | f_1, f_2, \dots, f_M) = \frac{e^{-\mu(k)} \mu(k)^n}{n!},$$

where $\mu(k)$ is defined as the expected number of peaks occurring within the k th bin. Using Eq. (1) we have

$$\mu(k) = \mu_C(k) + \sum_{m=1}^M \mu_{f_m}(k), \quad (2)$$

where $\mu_C(k)$ or $\mu_{f_m}(k)$ are defined as the integrals of the intensity functions $\rho_C(p)$ and $\rho(p|f_m)$ within bin k , respectively. We term these components the *rate functions* within bin k since they specify the expected number of peaks contributed by each note in bin k .

We assume that a single detection is made in a particular bin if one or more peaks are present on the continuous frequency scale. We then have the probability of a peak being detected in bin k as

$$P(N(k) \geq 1) = 1 - e^{-\mu(k)}$$

and the probability of no peak being detected is

$$P(N(k) = 0) = e^{-\mu(k)}.$$

Now, suppose that if a peak is detected at bin k we set observation $y_k = 1$, and $y_k = 0$ otherwise, $k = 0, \dots, K-1$. Hence for detected peak data $Y = [y_0, y_1, \dots, y_{K-1}]^T$ we have

$$P(Y|F) = \prod_{k=0}^{K-1} y_k (1 - e^{-\mu(k)}) + (1 - y_k) e^{-\mu(k)}. \quad (3)$$

Having obtained a likelihood function it is now in principle possible to perform inference on the number of notes and their pitches.

3. Rate function estimation

In order to compute the likelihood for a given note combination in Eq. (3) it is necessary to specify a rate function $\mu_f(k)$ for each possible note frequency f and for each frequency bin k . We note from Eq. (2) that the rate functions are expressed in terms of the underlying Poisson intensity functions. These intensity functions can in principle be learned from annotated training data. As an alternative, we here construct the rate functions $\mu_f(k)$ directly, either by learning their form from training data, or by construction from generic modeling principles:

Nonparametric estimation from training data. In this approach, rate functions are estimated from a large database of annotated training data. Peaks in the discrete Fourier transform (DFT) are extracted from each frame of data using a thresholded first-difference operation, with a frequency-dependent threshold determined using a running median filter. Their positions in terms of frequency bins are then histogrammed to determine the rate functions for each note separately.

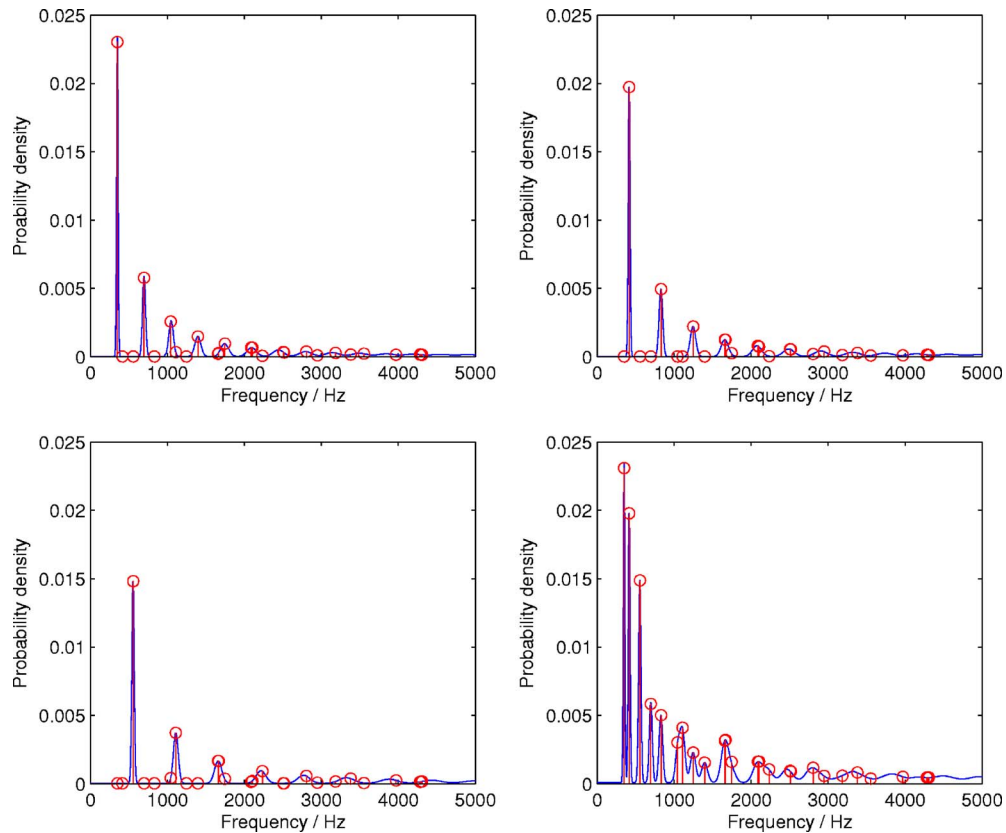


Fig. 2. (Color online) Rate function from a generic parametrized model.

Generic model. We may expect this to generalize better to a range of instruments. In this approach a Gaussian mixture model is proposed for the rate function. The mixture components are expected to be centered close to the frequency of the harmonic number h , i.e., at a frequency h_f . The general form of the rate function model for a note of fundamental frequency f is then as follows:

$$\mu_f(k) = \sum_{h=1}^H \frac{\beta_{f,h}}{\sqrt{2\pi\sigma_{f,h}^2}} \exp\left[-\frac{(f_k - hf)^2}{2\sigma_{f,h}^2}\right],$$

where f_k is the center frequency of bin k , $\sigma_{f,h}^2$ is the variance of that component's frequency, and $\beta_{f,h}$ are positive mixture weights (which need not sum to unity).

The variance and mixture weight components are constrained in a particular way such that $\beta_{f,h} = Ae^{-\beta h}$ and $\sigma_{f,h}^2 = \kappa^2 h^2$, where A , κ , and β are parameters to be specified or fitted to the peak data. See Fig. 2 for a realization of the rate function of a note using this model.

An alternative approach investigated was to estimate the parameters of the model from labeled training data. However, we found in practice that the performance of such a scheme was generally poorer than the model suggested above.

The clutter intensity ρ_C is modeled as uniform over the frequency range of the peak detector.

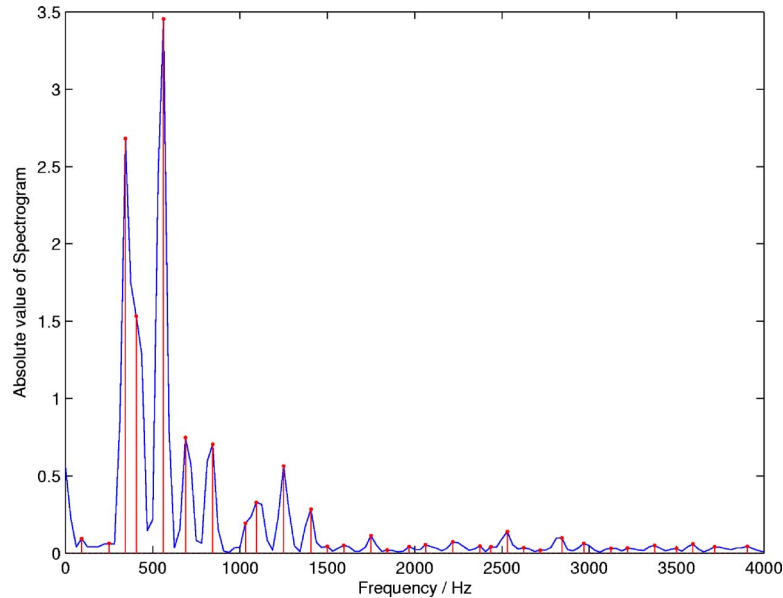


Fig. 3. (Color online) Fitting of a Poisson point process model (solid line) to peak position data (circles). The peaks correspond to three note pitches. Individual notes fitted shown in top left, top right, and bottom left panels, and the mixture of three notes fitted is shown in the lower-right panel.

4. Maximum-likelihood transcription

For this proof-of-principle implementation we perform a frame by frame maximum likelihood (ML) transcription of audio extracts. Prior to analysis, peaks are extracted as for the training data in the nonparametric approach in Sec. III. In addition a second differencing technique was found to detect some peaks that are otherwise obscured by nearby peaks with larger magnitude, detecting, for example, the peak at 400 Hz in Fig. 1 close to the peak at 350 Hz.

An exhaustive search of all the possible note combinations to find the ML solution is computationally infeasible for long extracts. Instead we iterate to find a local maximum. An effective approach was a greedy search algorithm that added at each iteration the note with greatest increase in likelihood. We verify this greedy search with a sampling procedure that takes a subset Q of m notes from the set $K = \{f_1, f_2, \dots, f_M\}$ of notes found, and checks that the ML solution of $(m+1)$ notes given Q is still a subset of K . We found that the greedy solution consistently passes this verification, and suggest this is due to the robust behavior of the Poisson likelihood function (3) over the search space, which renders each note in the mixture reasonably independent of the others. This suggestion requires further verification in more detailed studies. See Fig. 3 for an illustration.

5. Results

We demonstrate the two models on polyphonic, classical piano music, with up to four notes playing simultaneously. Frames were grouped into single “chord” entities using a time-frequency based segmentation procedure,¹³ with some manual intervention to correct for gross errors; and peaks from these grouped frames were analyzed together.

For cases where the number of simultaneous note pitches M is unknown, we estimate M by the Akaike information criterion (AIC).¹⁴ The AIC criterion is calculated as follows:

$$\text{AIC} = 2M' - 2 \ln p(Y|\widehat{\mu}_{M'}),$$

where $\widehat{\mu}_{M'}$ is the rate function corresponding to maximum likelihood estimate of M' simultaneous notes. We then choose M to be the value of M' for which AIC is a minimum.

Table 1. Performance of models on a set of piano music extracts. For the generic model, we have chosen the following set of parameters: $H=10, A=1, \kappa=1, \beta=0$.

Extract	Recording	M	Trained	Generic
Creation	Steinway	2	100%	100%
		3	90%	84%
		4	85%	78%
Moonlight	Elena Kuschnerova ^a	3–4		78%
	MIDI Synthesized ^b	3–4		91%
Variations	Andrew Koay ^c	4		88%

^aAvailable from www.elenakuschnerova.com

^bRecorded using Winamp (www.winamp.com)

^cAvailable from <http://music.download.com/>

The performance metric is $(N - M - E) / N$ where N is the correct number of notes from the ground truth, M is the number of notes missed from the ground truth, and E is the number of error notes not present in the ground truth.

Table 1 presents our results on the extracts (see Fig. 4) tested. Figure 5 demonstrates a transcription of the “Moonlight” extract. Results for all methods and extracts are very promising. The nonparametric trained model is observed to perform better than the generic model, but the training method used is not practical for many music transcription applications.

‘Creation’ - from *The Heavens are Telling* from Haydn’s *The Creation*

2-part (upper stave) & 3-part.



4-part



‘Moonlight’ - measures 1-8 of Beethoven’s Piano Sonata No. 14 ‘Moonlight’, 2nd. movement



‘Variations’ - measures 1-4 of Mendelssohn’s *Variations Sérieuses*



Fig. 4. Scores of extracts.

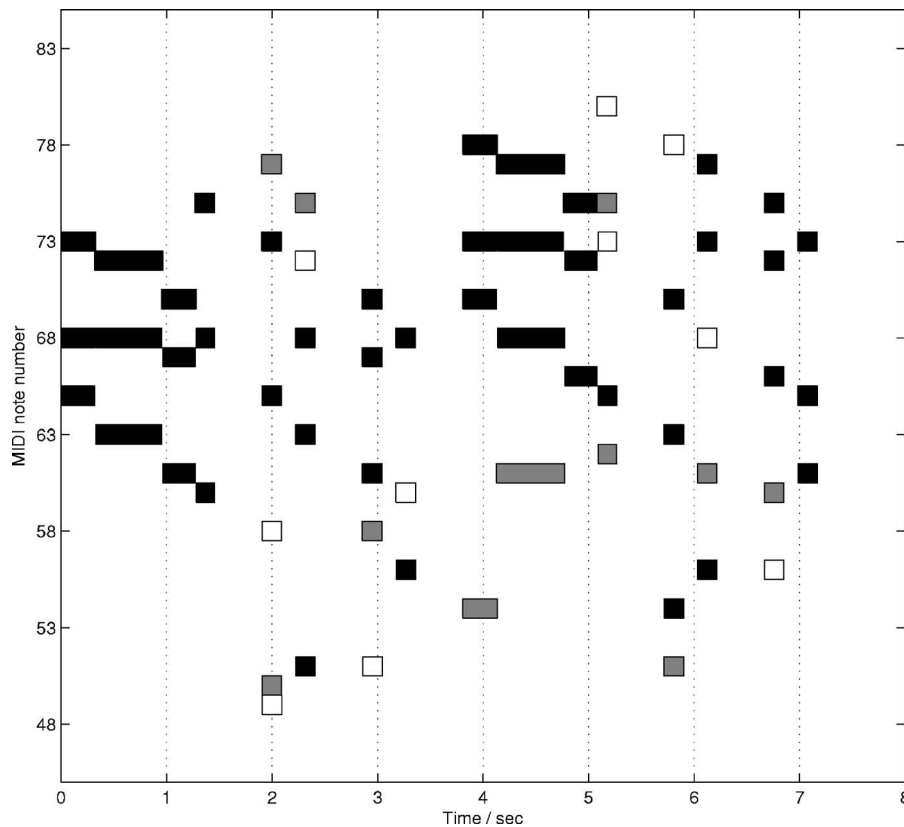


Fig. 5. Transcription of Moonlight using generic model and AIC: Black notes are correctly transcribed, white shows notes compared with the ground truth, and gray shows additional notes not present in the ground truth.

6. Conclusion and discussion

This paper has introduced Poisson point processes as a model for fundamental frequency estimation in polyphonic music. The principal advantage of such an approach is the simplicity of the resulting likelihood function, allowing many notes to be superimposed in a straightforward manner, and without performing any explicit data association task to link detected peaks with particular note harmonics. Several possible forms for the rate function were considered and example transcription results were given for polyphonic piano music.

We anticipate that performance gains can be achieved by embedding the Poisson model in a hierarchical model that links multiple frames together, thus directly modeling the evolution of pitches with time.¹⁵ A useful starting point is a hidden Markov model for pitch transitions over time, with a Poisson observation model for individual frames.

A further unexplored area is a model for the DFT amplitude process, which could guide the transcription process to better results and also lead to inference of additional quantities such as note timbre, playing volume, or instrument identity. Here one can consider extending the point process to a *marked point process*,⁸ in which both amplitudes and frequencies of peaks are modeled. Initial investigations have shown that this is a promising approach.

References and links

- ¹M. Davy and S. Godsill, "Bayesian harmonic models for musical pitch estimation and analysis," Technical Report No. CUED/F-INFENG/TR 431, Engineering Department, University of Cambridge, UK (2002).
- ²M. Davy and S. Godsill, "Bayesian harmonic models for musical signal analysis," in *Bayesian Statistics VII*, edited by J. Bernardo (Oxford University Press, Oxford) (2003).
- ³A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Speech*

- Audio Process. **14**, 679–694 (2006).
- ⁴A. P. Klapuri, “Multiple fundamental frequency estimation based on harmonicity and spectral smoothness,” *IEEE Trans. Speech Audio Process.* **6**, 804–816 (2003).
- ⁵H. Thornburg, R. Leistikow, and J. Berger, “Melody extraction and musical onset detection from framewise STFT data,” *IEEE Trans. Speech Audio Process.* (2006), accepted for publication.
- ⁶R. Maher, “Fundamental frequency estimation of musical signals using a two-way mismatch procedure,” *J. Acoust. Soc. Am.* **95**, 2254–2263 (1994).
- ⁷J. Bello, L. Daudet, and M. Sandler, “Automatic piano transcription using frequency and time-domain information,” *IEEE Trans. Speech Audio Process.* **14**, 2242–2251 (2006).
- ⁸D. R. Cox and V. Isham, *Point Processes* (Chapman and Hall, London, 1980).
- ⁹R. J. Leistikow, H. Thornburg, J. Smith III, and J. Berger, “Bayesian identification of closely-spaced chords from single-frame STFT peaks,” in *Proceedings of the 7th International Conference on Digital Audio Effects*, Naples, Italy (2004).
- ¹⁰K. Gilholm and D. Salmond, “Extended object and group tracking,” in *RTO SET Symposium on Target Tracking and Sensor Data Fusion for Military Observation Systems* (2003).
- ¹¹K. Gilholm and D. Salmond, “A spatial distribution model for tracking extended objects,” *IEE Proc., Radar Sonar Navig.* **152**, 364–371 (2005).
- ¹²K. Gilholm, S. Godsill, S. Maskell, and D. Salmond, “Poisson models for extended target and group tracking,” in *SPIE Conference 2005: Signal and Data Processing of Small Targets* (2005).
- ¹³H. Nagano, K. Kashino, and H. Murase, “A fast search algorithm for background music signals based on the search for numerous small signal components,” *Proc. of the 2003 International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, Vol. 5, 796–799 (2003).
- ¹⁴H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Comput.-Aided Des.* **19**, 716–723 (1974).
- ¹⁵S. Godsill, “Computational modeling of musical signals,” *CHANCE Magazine* **17** (2004).