Dynamic Bayesian networks for symbolic polyphonic pitch modeling

Stanislaw Raczynski, Emmanuel Vincent, Shigeki Sagayama

Dynamic Bayesian networks for symbolic polyphonic pitch modeling

Stanisław A. Raczyński, Emmanuel Vincent, Shigeki Sagayama

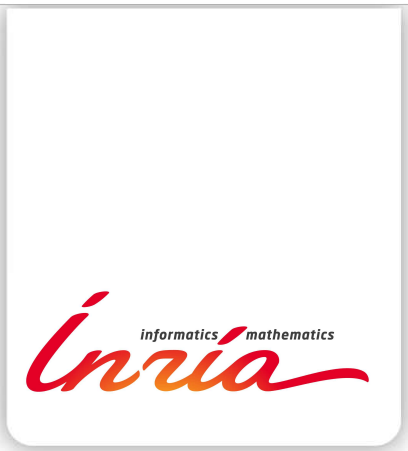
TECHNICAL REPORT

N° 430

September 2012

Project-Team METISS

ISSN 0249-0803 ISRN INRIA/RT--430--FR+ENG



Dynamic Bayesian networks for symbolic polyphonic pitch modeling

Stanisław A. Raczyński, Emmanuel Vincent, Shigeki Sagayama

Project-Team METISS

Technical Report n° 430 — September 2012 — 28 pages

Abstract: Symbolic pitch modeling is a way of incorporating knowledge about relations between pitches into the process of analyzing musical information or signals, and it is typically done in a statistical framework. It has proven to be an efficient way of improving the performance of various Music Information Retrieval (MIR) algorithms. In this paper, we propose a family of probabilistic symbolic polyphonic pitch models for multiple pitch estimation, which account for both the "horizontal" and the "vertical" pitch structure. These models are formulated as linear or log-linear interpolations of up to five submodels, each of which is responsible for modeling a different aspect of music.
The ability of the models to predict symbolic pitch data is evaluated in terms of their cross-entropy, and of a newly proposed "contextual cross-entropy" measure. Their performance is then measured on synthesized polyphonic audio signals in terms of the accuracy of multiple pitch estimation in combination with a Nonnegative Matrix Factorization-based acoustic model. In both experiments, the log-linear combinations of at least one "horizontal" (e.g. harmony) and one "vertical" (e.g. note duration) models outperformed the baseline methods, by almost 60% in cross-entropy reduction and almost 4% in multiple pitch estimation accuracy. This work provides a proof of concept of the usefulness of model interpolation in the area of pitch modeling, which may be used for improved symbolic modeling in the future.

Key-words: Dynamic Bayesian Networks, multipitch analysis, symbolic pitch modeling

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Réseaux bayésiens dynamiques pour l'estimation de hauteur polyphonique

Résumé : La modélisation symbolique de la hauteur est une façon d'incorporer les connaissances sur les liens entre hauteurs dans l'analyse de données ou de signaux musicaux et elle s'effectue typiquement dans un cadre statistique. Elle s'est révélée être une façon efficace d'améliorer la performance de divers algorithmes de *Music Information Retrieval* (MIR). Dans ce papier, nous proposons une famille de modèles symboliques de la hauteur pour l'estimation de hauteurs multiples qui rend compte à la fois de l'agencement "horizontal" et "vertical" des hauteurs. Ces modèles sont formulés comme le résultat de l'interpolation linéaire ou log-linéaire de deux à cinq sous-modèles représentant chacun un aspect différent de la musique. La capacité de ces modèles à prédire des hauteurs symboliques est évaluée en terme d'entropie croisée et d'une nouvelle mesure d'"entropie croisée contextuelle" que nous proposons. Leur performance est ensuite mesurée en combinaison avec un modèle acoustique basé sur la factorisation en matrices positives pour l'estimation de hauteurs multiples sur des signaux audio polyphoniques synthétisés. Dans ces deux expériences, la combinaison log-linéaire d'au moins un modèle "horizontal" (par exemple de l'harmonie) et un modèle "vertical" (par exemple de la durée des notes) dépasse la performance des méthodes de référence d'au moins 60% en terme de réduction d'entropie croisée et 4% en terme de précision d'estimation de hauteurs multiples. Ce travail fournit une preuve de concept de l'utilité de l'interpolation de modèles pour la modélisation de la hauteur, qui pourra être utilisée à l'avenir pour une meilleure modélisation de la musique symbolique.

Mots-clés : Réseaux bayésiens dynamiques, analyse de la hauteur polyphonique, modélisation symbolique de la hauteur

1 Introduction

Symbolic music modeling, also known as *musicological modeling* [18, 19, 25], is the equivalent of language modeling in speech processing. It has the potential to improve the performance of many Music Information Retrieval (MIR) tasks, such as multiple pitch estimation [25], algorithmic composition [6, 27] and automatic performance [5, 11], chord and key estimation [19, 29, 17], music structure analysis [18], etc., as a part of an integrated statistical model of music [31].

A particular MIR task, polyphonic pitch transcription, consists of the estimation of the pitches, the onset times and the durations of each of the musical notes present in a recorded audio signal. Many techniques have been proposed to deal with the pitch transcription: sparse coding [1], auditory filterbanks [25, 14], optimal harmonic amplitude summation [13] or spectrotemporally constrained Gaussian mixture models [9], but the most popular methods are based on Nonnegative Matrix Factorizations (NMF) and its variations [4, 21, 28, 30, 16, 2]. Except for [9], all of those solutions work in two subsequent steps (though much of the work focuses only on the first one): first, the confidence of the presence of a pitch is quantified for every spectrotemporal bin by the acoustic model (sparse coder, filterbank, NMF). The confidences are then post-processed and grouped to form the detected musical notes. Without including any prior knowledge about the occurances of the notes (a symbolic pitch model) $P(\mathbf{N})$, this approach can be considered as a form of maximum likelihood estimation:

$$\widehat{\mathbf{N}} = \arg\max_{\mathbf{N}} \mathrm{P}(\mathbf{S}|\mathbf{N}), \tag{1}$$

where $\mathrm{P}(\mathbf{S}|\mathbf{N})$ is the acoustic model. Adding a note prior would mean result in an estimation of the notes in the maximum a posteriori-like manner:

$$\widehat{\mathbf{N}} = \arg\max_{\mathbf{N}} \mathrm{P}(\mathbf{S}|\mathbf{N})\mathrm{P}(\mathbf{N}). \tag{2}$$

While acoustic modeling has been widely studied, symbolic pitch modeling has been given much less attention so far. Some researchers have used basic musicological models in order to overcome the limitations of current state-of-the-art multiple pitch transcription models: Ryynänen and Klapuri proposed in [26] a melody transcription method that uses a Hidden Markov Model (HMM) to model note envelopes, together with a simple musical key model, but their approach was limited to monophonic note sequences. A polyphonic extension was later proposed in [25], but that model still lacks modeling of the dependencies between concurrent pitches: the music is treated as a combination of independent and non-overlapping melodic voices. In other MIR areas, Raphael and Stoddard have [24] proposed to use an HMM as a symbolic model for harmonic analysis, i.e., for the estimation of the chord progression behind a sequence of notes. Similar HMMs have also been successfully used for harmonic analysis of audio signals (for a recent paper see e.g. [29]). These HMM-based approaches, however, model only chromatic pitch classes instead of actual absolute pitches, and the temporal dependencies are only present between chords.

We propose a family of probabilistic pitch models based on Dynamic Bayesian Networks (DBNs), which account both for the "vertical" dependencies between concurrent notes (harmony) and the "horizontal" dependencies between notes and chords. The main challenge when building such a model is dealing with the high dimensionality of the resulting distributions that makes the training and inference very difficult or even impossible in practice. In our previous work, [23], we have dealt with this by applying a series of factorizations and approximations to the conditional note combination distribution $P(\mathbf{N}_t|\mathbf{N}_{1:t-1}, C_t)$ and the inference was performed on a highly reduced solution space. However, that was still problematic because that distribution could not be normalized over the entire solution space and the result was not a true probabilistic model, while the normalization was very computationally expensive. In this paper, we propose to factorize the note combination distribution into a product of single note note distributions, each modeled with multiple, easy to normalize, conditional Bernoulli models that are combined by means of linear or log-linear interpolation.

This paper is organized as follows: section 2 details the proposed approach and the way of combining them through interpolation. Particular distributions chosen in this work are discussed in section 3. Section 4 describes then the experimental set-up and the results of symbolic and audio evaluations. Finally, the conclusion is given in section 5.

2 General approach

2.1 DBN structure

We model the prior distribution of the note sequences $\mathrm{P}(\mathbf{N})$ as a DBN with two layers of nodes: a chord (harmony) layer $\mathbf{C} = (C_1, C_2, \ldots, C_T)$ and a note activity layer $\mathbf{N} = (\mathbf{N}_1, \mathbf{N}_2, \ldots, \mathbf{N}_T)$, where T is the number of time frames in the analyzed note sequence. We then assume that these sequences are first-order Markovian [29]:

$$\mathrm{P}(\mathbf{N}) = \sum_{\mathbf{C}} \mathrm{P}(C_1)\mathrm{P}(\mathbf{N}_1|C_1) \prod_{t=2}^{T} \mathrm{P}(\mathbf{N}_t|\mathbf{N}_{t-1}, C_t)\mathrm{P}(C_t|C_{t-1}). \tag{3}$$

$\mathbf{N}_t$ is a set of boolean variables encoding the activity of notes in the time frame t on the discrete MIDI pitch scale. The DBN structure corresponding to the avove equation is presented in Fig. 1.

2.2 Interpolation

Unfortunately, the note activity probability distribution $\mathrm{P}(\mathbf{N}_t|\mathbf{N}_{t-1}, C_t)$ is a highly-dimensional discrete distribution, too complex to be trained or used for inference in practice, as it would require $2^{88} \times 2^{88} \times 24$ distinct probability values to be defined in the case of a full piano scale for example. To deal with this

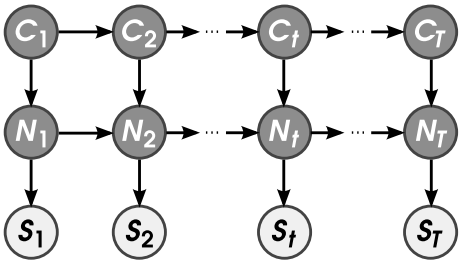


Figure 1: Proposed Dynamic Bayesian Network structure for polyphonic pitch modeling.

problem, we first factorize the distribution by applying Bayes rule:

$$\mathrm{P}(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}) = \prod_{k=1}^{K} \mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1}), \tag{4}$$

where k is the analyzed pitch and K is the size of the analyzed pitch range. We then approximate this distribution using a combination of several *submodels*, combined by means of *linear interpolation*

$$\mathrm{P}(N_{t,k}|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) \approx Z^{-1} \sum_i \lambda_i \mathrm{P}_i(N_{t,k}|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1}), \tag{5}$$

or *log-linear interpolation*

$$\mathrm{P}(N_{t,k}|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) \approx Z^{-1} \prod_i \mathrm{P}_i(N_{t,k}|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1})^{\lambda_i}, \tag{6}$$

where $\lambda = \{\lambda_i\}$ are interpolation coefficients, P_i are submodels and Z is the normalization factor:

$$Z = \sum_{l=0}^{1} \sum_i \lambda_i \mathrm{P}_i(N_{t,k} = l|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1}) \tag{7}$$

for the linear interpolation and

$$Z = \sum_{l=0}^{1} \prod_i \mathrm{P}_i(N_{t,k} = l|\mathbf{N}_{t-1}, C_t, \mathbf{N}_{t,1:k-1})^{\lambda_i} \tag{8}$$

for the log-linear one. Each submodel is responsible for modeling a different musicological aspect of the note sequences and can therefore be defined over a subset of the conditioning variable set.

Linear interpolation of models have been first proposed in the context of spoken language modeling by Jelinek and Mercer in [8], while log-linear interpolation was proposed much later by Klakow [12]. In that work, however, the submodels are univariate models with temporal dependencies only. In pitch modeling, we extend the concept of model interpolation to arbitrary dependencies.

3 Considered submodels

In this work, we define 5 submodels as a proof of concept: the chord model and harmony submodel are responsible for modeling chord progressions and relations between chords and pitches; the note duration submodel deals with note and silence durations, and transitions between them for individual pitches; the voice movement submodel models melodic intervals in voices; the neighbor submodel handles relations between vertically neighboring pitches; and finally the polyphony submodel models the degree of polyphony in each time frame. Other submodels are naturally possible, but we believe that the above set covers all the aspects of musical knowledge that are important for multiple pitch analysis.

We will now describe each of these submodels in detail and show the corresponding probabilities, as trained on the data described in section 4.

3.1 Chord model

The chord transition probability $\mathrm{P}(C_t|C_{t-1})$ is easy to model with a multinomial (categorical) probability distribution. This approach is common in MIR tasks that deal with chord progression, e.g. in chord recognition [29]. It is also common to assume a 24-word chord dictionary, i.e., 12 major and 12 minor chords. We have adopted this approach as well, so the chord transition distribution is described in terms of a 24×24 transition matrix.

The left part of Fig. 2 shows the chord transition matrix trained on the entire available dataset. Unfortunately, the obtained transition probabilities are biased, as some keys, and therefore some chord progressions, are sparsely represented in our dataset, while others dominate. However, we can assume that the chord transitions have the same distribution in all keys if observed in relation to the tonic, which is reasonable since any song can be transposed to an arbitrary key without any loss in musical correctness. In other words, we assume that the same probability should be given to e.g. the transition from C-major to F-major chord (I→IV transition in C-major key) and the transition from A♭-major to D♭-major (I→IV transition in A♭-major key). In this case, the chord transition probability is a function only of the interval between chord roots and their types. The transition matrix obtained by tying distributions in the above way is presented in Fig. 2.

Furthermore, because key is not considered in our model, we assume a uniform distribution of the initial chord $\mathrm{P}(C_1) = \text{const.}$, which in classical Western music is always the tonic.

3.2 Harmony submodel

Similarly to the chord model, in order to avoid overfitting, we tie together the probabilities of notes having the same musicological function. This, again, is based on the observation that music can be freely transposed between keys and so the notes should have identical distribution with respect to the chord's root

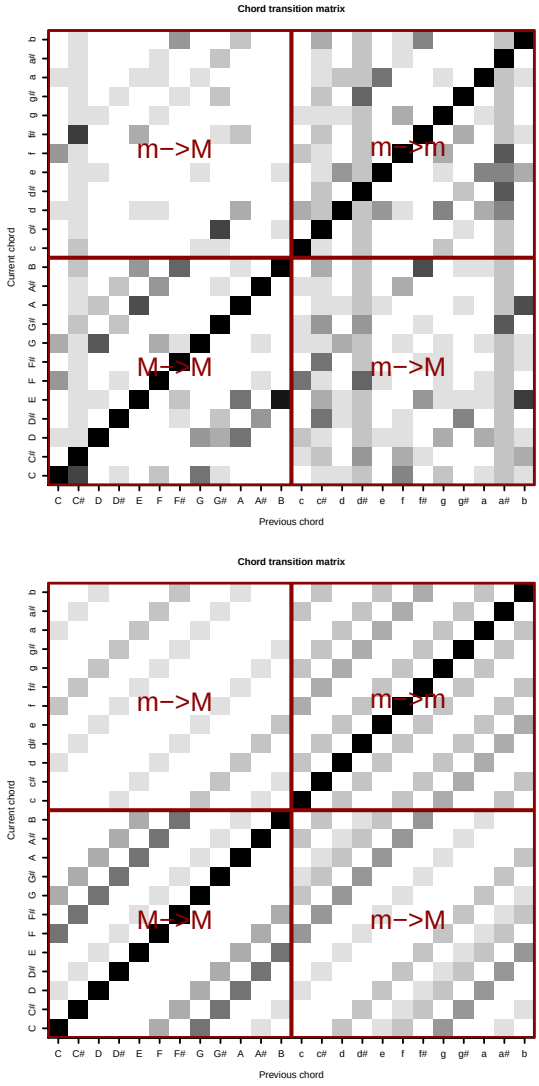


Figure 2: Chord transition probability matrix if state tying was not used (top) and if transition probabilities were tied (bottom). Darker color represent higher probability values. Minor chords are annotated with lower case (m) and major chords with upper case (M).

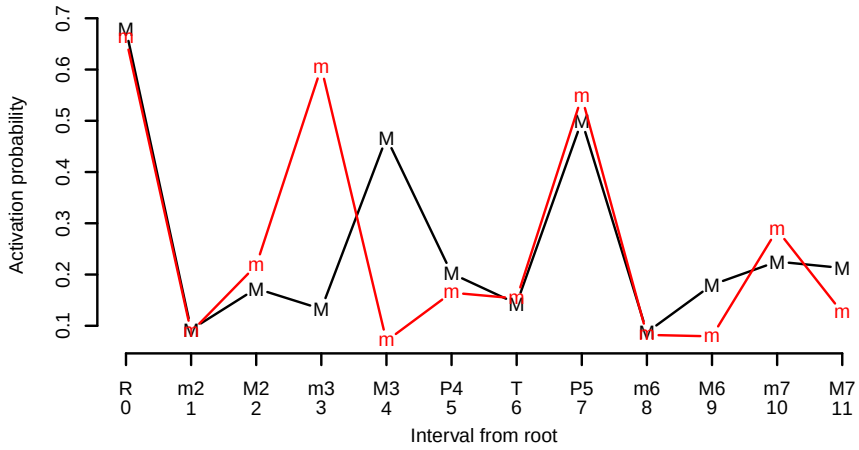


Figure 3: Pitch probability distribution for major (M) and minor (m) chords as a function of the interval from the chord's root note.

notes.

$$\mathrm{P}_1(\mathbf{N}_t|C_t,\mathbf{N}_{t-1},\mathbf{N}_{t,1:k-1}) = \mathrm{P}(\,\mathrm{inter}\{k;\mathrm{root}\{C_t\}\}|\mathrm{mode}\{C_t\}), \tag{9}$$

where "inter" is the musical interval operator, "root" is the root note operator and "mode" is the mode operator, i.e., major or minor. The corresponding probability distribution is presented in Fig. 3.

3.3 Duration submodel

In this submodel the individual note activities are assumed to be dependent only on the previous state of the same pitch:

$$\mathrm{P}_2(\mathbf{N}_t|C_t,\mathbf{N}_{t-1},\mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|N_{t-1,k}). \tag{10}$$

This in fact is a pitch-dependent conditional Bernoulli model. Its parameters are presented in Fig. 4.

3.4 Voice submodel

In this submodel, we assume than the note activity depends only on the distance to the closest active pitch in the previous frame:

$$\mathrm{P}_3(\mathbf{N}_t|C_t,\mathbf{N}_{t-1},\mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|M_{t,k}), \tag{11}$$

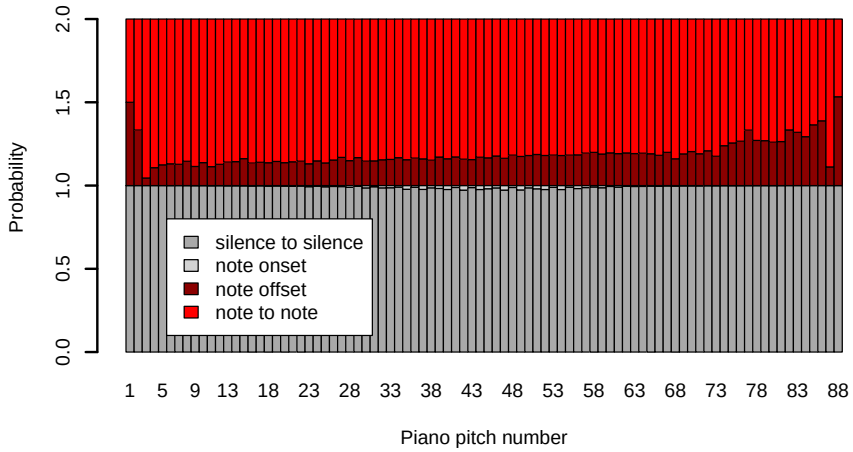


Figure 4: The duration submodel $\mathrm{P}(N_{t,k}|N_{t-1,k})$.

where $M_{t,k} = |k - j|$ is the distance between the given pitch k and the closest active pitch in the previous time frame j. If there was no pitch in the previous time frame, then $M_{t,k} = 88$. If the pitch k was active in the previous time frame, this model acts as a duration model, otherwise it is a simple voice movement model. Trained parameters for this submodel are depicted in Fig. 5.

3.5 Polyphony submodel

The polyphony submodel models the number of notes active simultaneously:

$$\mathrm{P}_4(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|L_{t,k}), \tag{12}$$

where $L_{t,k} = \sum_{m=1}^{k-1} N_{t,m}$. The resulting distribution is plotted in Fig. 6.

3.6 Neighbor submodel

This submodel captures the note probability given the note activities directly below it:

$$\mathrm{P}_5(\mathbf{N}_t|C_t, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}) = \mathrm{P}(N_{t,k}|N_{t,k-1}, N_{t,k-2}). \tag{13}$$

It is a binary trigram model. Its parameters are presented in Fig. 7.

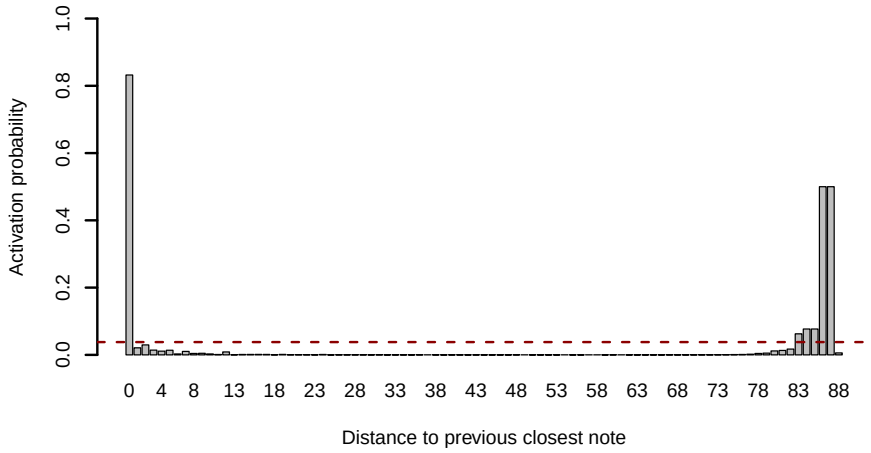


Figure 5: The voice submodel $\mathrm{P}(N_{t,k} = 1|M_{t,k})$. As the distance increases, the probabilities quickly decrease, with peaks at, e.g., the perfect fifth and the octave, then increase again as the training data sparsity increases, tending to a uniform distribution at $M_{t,k} = 86$ and 87.

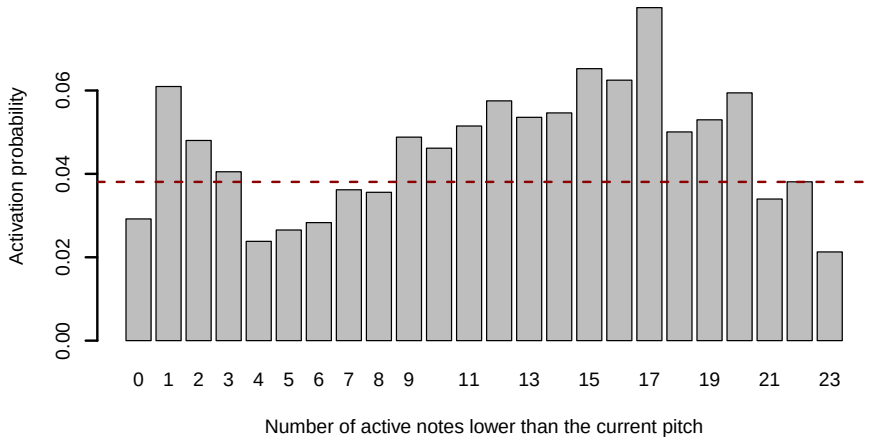


Figure 6: The polyphony submodel $\mathrm{P}(N_{t,k} = 1|L_{t,k})$. The dashed line marks the marginal note activity probability $\mathrm{P}(N_{t,k})$.

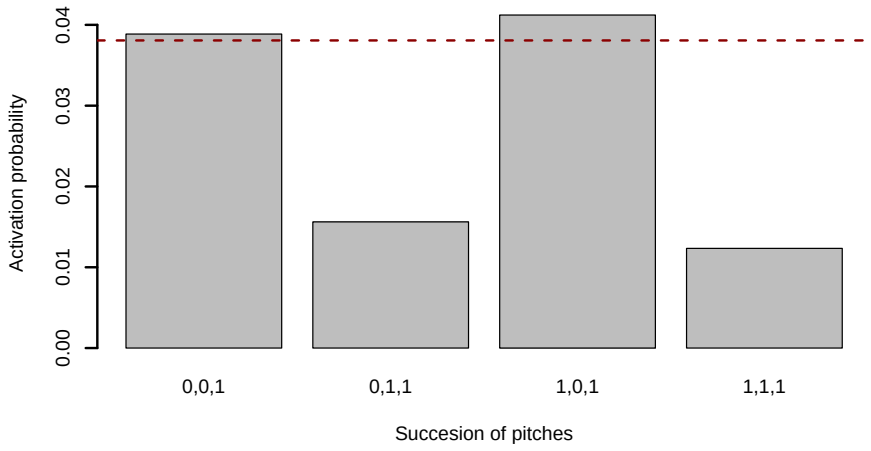


Figure 7: Neighbor model $\mathrm{P}(N_{t,k}|N_{t,k-1}, N_{t,k-2})$. The dashed line marks the marginal note activity probability $\mathrm{P}(N_{t,k})$.

4 Experiments

4.1 Data

Two datasets were used in the experiments: the widely used RWC database [7] and the Mutopia Project dataset [20]. The classical pieces of the RWC database were annotated with detailed harmony labels that include: keys and modulations, and chords with their roots, inversions, types and various modifications [10]. This data uses abstract, tempo-independent musical time (measures and beats), and served as the chord ground-truth for training the harmony and chord models.

The Mutopia dataset consisted of 1468 files divided into 3 subsets: for training (1268 files), validation (100 files) and testing (100 files). The training set was used to train all remaining submodels, while the smoothing parameters and the interpolation weights from Eqs 5 and 6 were trained on the validation set. Experiments were performed on the testing data. All symbolic data was score-like (as opposed to real performance data) and time-quantized so that 1 frame corresponded to $1/6^{\mathrm{th}}$ of a beat.

The Mutopia dataset contains music played on a variety of instruments: chordophones (piano, guitar, cello, shamisen, violin, viola), aerophones (church, rock and reed organs, clarinet, oboe, French horn, bassoon, pan flute, recorder, trumpet), as well as singing voice (in chorus).

4.2 Smoothing

In all cases a simple, additive smoothing was used, with the smoothing parameter α optimized for each model separately to maximize its log-likelihood on the validation dataset.

4.3 Interpolation coefficients

The interpolation weights $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ from Eqs 5 and 6 are optimized by maximizing their log-likelihood:

$$\widehat{\lambda} = \arg\max_{\lambda} \log \mathrm{P}(\mathbf{N}|\lambda), \tag{14}$$

calculated on the validation dataset. Optimization was performed using a non-negatively constrained limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) method (a quasi-Newton optimization), built into the GNU R environment as the `optim()` function. The resulting values are listed in Tables 1 and 2. The initial coefficient values were all set to $\lambda_p = 1$, $p \in \{1, \ldots, 5\}$.

Coefficient	**Model**	**DN**	**HCV**	**HCDPN**	**HCDVPN**
λ_1	Harmony	—	0.939	0.896	0.907
λ_2	Duration	0.980	—	0.863	0.272
λ_3	Voice	—	0.847	—	0.570
λ_4	Polyphony	—	—	0.000	0.000
λ_5	Neighbor	0.024	—	0.000	0.000

Table 1: Trained interpolation coefficients for different combinations of the submodels, obtained for the log-linear interpolation.

Individual models were not given coefficients for the linear interpolation, because the model normalization always forces them to 1, but they were applied for the log-linear interpolation and acted as a sort of fudge factors.

Coefficient	**Model**	**DN**	**HCV**	**HCDPN**	**HCDVPN**
λ_1	Harmony	—	0.000	0.000	0.000
λ_2	Duration	1.000	—	1.000	0.376
λ_3	Voice	—	1.000	—	0.623
λ_4	Polyphony	—	—	0.000	0.000
λ_5	Neighbor	0.000	—	0.000	0.000

Table 2: Trained interpolation coefficients for different combinations of the submodels, obtained for the linear interpolation.

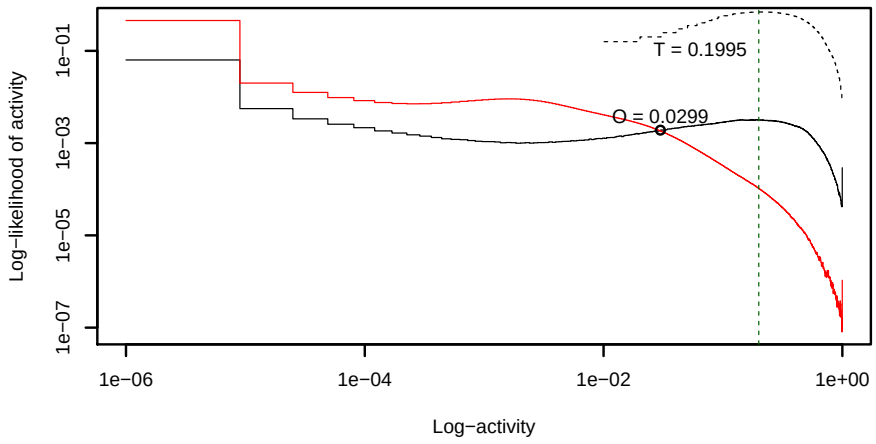


Figure 8: Note salience distributions: $\mathrm{P}(S_{t,k}|N_{t,k} = 1)$ (black line) and $\mathrm{P}(S_{t,k}|N_{t,k} = 0)$ (red line). The dashed line shows the average F-measure resulting from applying different threshold values to the test dataset, for reference.

We see from the resulting coefficient values, that the Polyphony and Neighbor submodels received very low, or even zero weights in all the cases, which means that either the information they hold overlaps with other used models, or that they were not able to capture much useful information about the notes. Typically, only one "vertical" submodel would get a non-zero weight and the two "horizontal" submodels (Duration and Voice) would share a similar non-zero coefficient value, which again suggests some information overlap between the vertical submodels.

4.4 Salience model

The observed note saliences are assumed to be iid. given the note activities:

$$\mathrm{P}(\mathbf{S}_t|\mathbf{N}_t) = \prod_{k=1}^{88} \mathrm{P}(S_{t,k}|N_{t,k}). \tag{15}$$

Both $\mathrm{P}(S_{t,k}|N_{t,k} = 0)$ and $\mathrm{P}(S_{t,k}|N_{t,k} = 1)$ were estimated by measuring histograms of the detected salience. Before calculating the histograms, the saliences were nonlinearily transformed by applying an exponential factor $\chi = 0.5$ in order to enhance estimation precision for the low salience values. Number of histogram bins was set to 500. The obtained salience distributions are presented in Fig. 8.

4.5 Symbolic evaluation

The models are compared by calculating the cross-entropy of the observed note data given the musicological prior:

$$\mathrm{H}(\mathbf{N}) = -\frac{1}{88T}\log_2 \mathrm{P}(\mathbf{N}|\Lambda) \tag{16}$$

where Λ is the musicological model, T is the number of frames ($88T$ is the number of all analyzed note activities, used as a normalizing constant to obtain an average cross-entropy over all the time-pitch bins).

4.5.1 Para-cross-entropy

However, comparing models of different structure using the regular cross-entropy turned out to be difficult (see Figs 10 and 11), as the values are biased by the abundance of silence in the activity matrices. It is therefore beneficial to observe the cross-entropy only on selected parts of those matrices. Thus, if the harmony model is not used, we can calculate the cross-entropy as

$$\mathrm{H} = -\frac{1}{88T}\log_2 \prod_{t=1}^{T}\prod_{k=1}^{88} \mathrm{P}(N_{t,k}|\mathbf{N}_{1:t-1}, N_{t,1:k-1}), \tag{17}$$

$$\mathrm{H} = -\frac{1}{88T}\sum_{t=1}^{T}\sum_{k=1}^{88} \log_2 \mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}, N_{t,1:k-1}). \tag{18}$$

The averaging in the equation above can now be done over specific bins: active bins (notes), inactive bins (silence), onset bins or offset bins only:

$$\mathrm{pH}(S) = \frac{1}{|S|}\sum_{(t,k)\in S} -\log_2 \mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}, N_{t,1:k-1}), \tag{19}$$

where S is a subset of all the bins. We will refer to this measure as the *para-cross-entropy*.

If the harmony model is used however, we need to integrate over all possible chord sequences:

$$\mathrm{H}(\mathbf{N}) = -\frac{1}{88T}\log_2 \sum_{\mathbf{C}} \mathrm{P}(\mathbf{N}|\mathbf{C},\Lambda)\mathrm{P}(\mathbf{C}|\Lambda), \tag{20}$$

This integration is done with the Forward/Backward algorithm. The forward probability vector $\mathbf{f}_t$ for frame t is defined as the joint distribution of note all notes observed up to the current time frame and the chord value C_t at time t:

$$f_{t,i} = \mathrm{P}(\mathbf{N}_{1:t}, C_t = i), \tag{21}$$

where $i \in \{1, 2, \ldots, 24\}$. Its normalized form $\hat{\mathbf{f}}_t$ is the chord distribution given all previously observed notes:

$$\hat{f}_{t,i} = \mathrm{P}(C_t = i|\mathbf{N}_{1:t}), \tag{22}$$

Let us now notate the chord transition probability as $A_{i,j} = \mathrm{P}(C_t = i|C_{t-1} = j)$ and the note posterior as

$$d_{t,i} = \mathrm{P}(\mathbf{N}_t|C_t = i, \mathbf{N}_{t-1}) = \prod_{k=1}^{88} \mathrm{P}(N_{t,k}|C_t = i, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1}). \tag{23}$$

The forward vectors are calculated as

$$\hat{\mathbf{f}}_1 = p_1^{-1}\mathbf{d}_t \odot \pi, \tag{24}$$

$$\hat{\mathbf{f}}_t = p_t^{-1}\mathbf{d}_t \odot \hat{\mathbf{f}}_{t-1}\mathbf{A}, \tag{25}$$

where p_t is the normalizing factor:

$$p_t = \sum_{i=1}^{24} \mathrm{P}(\mathbf{N}_t, C_t = i|\mathbf{N}_{1:t-1}) = \mathrm{P}(\mathbf{N}_{1:t}|\mathbf{N}_{1:t-1}). \tag{26}$$

Because $\prod_{t=1}^{T} p_t = P(\mathbf{N}_{1:T})$, the normalizing factors can be used to calculate the cross-entropy:

$$\mathrm{H} = -\frac{1}{88T}\log_2 \prod_{t=1}^{T} p_t. \tag{27}$$

With the above formulation, we can track the cross-entropy over a subset of bins S. Let us define the following probabilities:

$$\mathring{h}_{t,i} = \mathrm{P}(\mathbf{N}_{t,k\notin S}, C_t = i|\mathbf{N}_{1:t-1}), \tag{28}$$

$$r_t = \sum_{i=1}^{24} \mathring{h}_{t,i} = \mathrm{P}(\mathbf{N}_{t,k\notin S}|\mathbf{N}_{1:t-1}), \tag{29}$$

$$\hat{\mathring{h}}_{t,i} = r_t^{-1}\mathring{h}_{t,i} = \mathrm{P}(C_t = i|\mathbf{N}_{1:t-1}, \mathbf{N}_{t,k\notin S}), \tag{30}$$

$$\dot{h}_{t,i} = \mathrm{P}(\mathbf{N}_{t,k\in S}, C_t = i|\mathbf{N}_{1:t-1}, \mathbf{N}_{t,k\notin S}), \tag{31}$$

$$q_t = \sum_{i=1}^{24} \dot{h}_{t,i} = \mathrm{P}(\mathbf{N}_{t,k\in S}|\mathbf{N}_{1:t-1}, \mathbf{N}_{t,k\notin S}). \tag{32}$$

q_t can be obtained from the forward vectors:

$$\hat{\mathring{\mathbf{h}}}_t = r_1^{-1}\mathring{\mathbf{d}}_t \odot \hat{\mathbf{f}}_{t-1}\mathbf{A}, \tag{33}$$

$$q_t = |\dot{\mathbf{d}}_t \odot \hat{\mathring{\mathbf{h}}}_t|, \tag{34}$$

where $\mathring{\mathbf{d}}_t = \prod_{k\notin S} \mathrm{P}(N_{t,k}|C_t = i, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1})$ and $\dot{\mathbf{d}}_t = \prod_{k\in S} \mathrm{P}(N_{t,k}|C_t = i, \mathbf{N}_{t-1}, \mathbf{N}_{t,1:k-1})$. Finally, the para-cross-entropy is obtained as a product of q_ts:

$$\mathrm{pH}(S) = -\frac{1}{|S|}\log_2 \prod_{t\in S} q_t \tag{35}$$

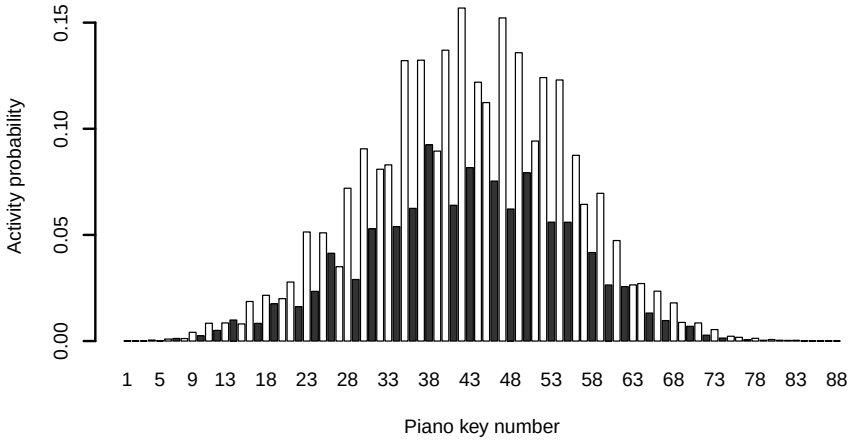


Figure 9: Parameters p_k of the independent note activity model. Black and white bars correspond to black and white piano keys, respectively.

4.5.2 Results

In the experiments, the para-cross-entropies were obtained for:

- individual note models: harmony (H), harmony + chord (HC), duration (D), voice (V), polyphony (P) and neighbor (N) model,
- model tandems: duration + neighbor (DN) and harmony + chord + voice (HCV) models,
- multiple models: HCDPV and HCDVPN,
- two reference models for comparison: an independent and identically distributed Bernoulli model $\mathrm{P}(N_{t,k}) \sim \mathrm{Bernoulli}(p)$ with $p = 0.03807$ and an independent, pitch-dependent Bernoulli model $\mathrm{P}(N_{t,k}) \sim \mathrm{Bernoulli}(p_k)$. The values of p_k are shown in Fig. 9.

Models were combined using both the linear and log-linear interpolation. The resulting values are presented in Figs 10 and 11. Comparing the regular cross-entropy with the para-cross-entropy for silence, we see how much the latter dominates the former and why is it better to use para-cross-entropy calculated for notes, onsets or offsets alone. Also, looking at the onset para-cross-entropies, we can observe the importance of using the harmony submodel: even though the regular cross-entropy is big for that model, it works very well at the relatively rare, but very important time of note onsets, when the other models fail to capture much information about the notes.

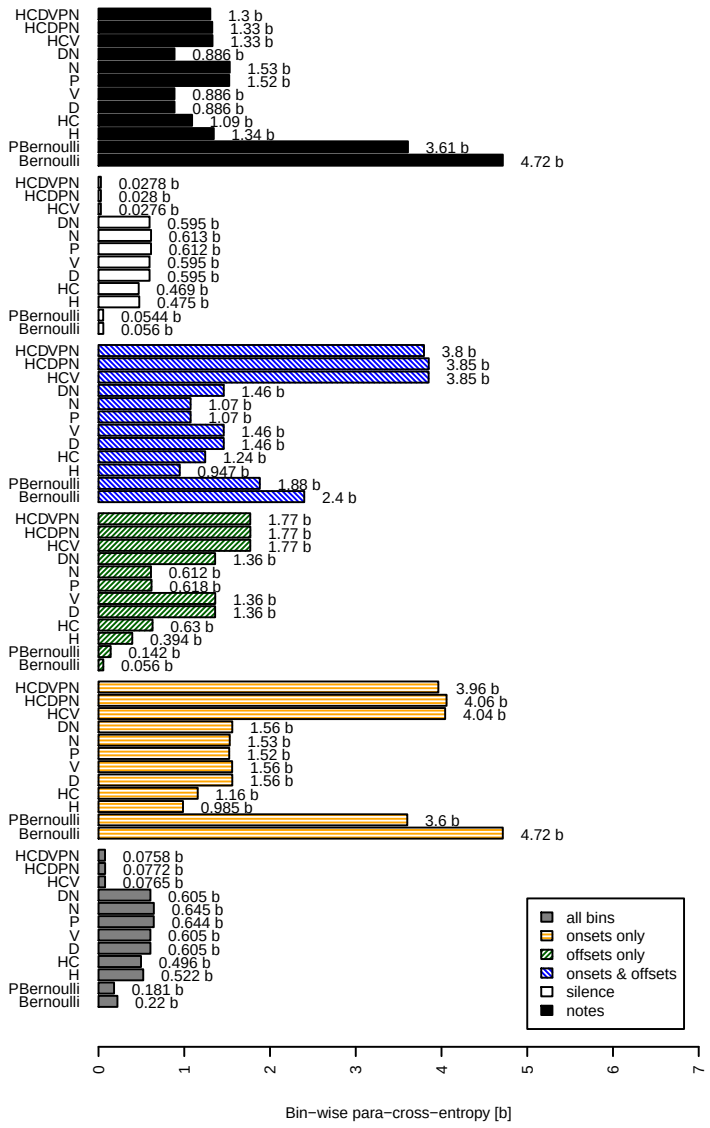


Figure 10: Para-cross-entropies calculated on evaluation dataset for linear interpolation.

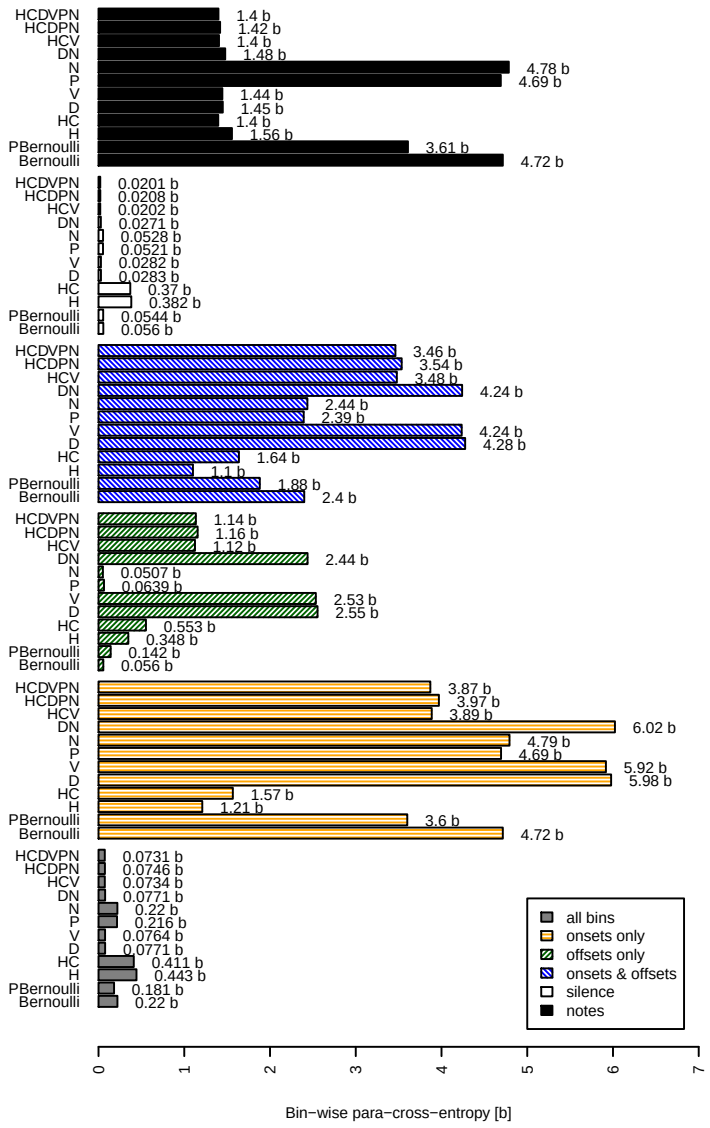


Figure 11: Para-cross-entropies calculated on evaluation dataset for log-linear interpolation.

4.6 Audio signal analysis

In the second part of the experimentation, we have used the developed models to perform multiple pitch estimation on audio signals. Decoding the most likely sequence of notes was performed with a frontier algorithm proposed in [15], modified to find the most likely path through all the hidden states: the integration (summation) was exchanged with finding the maximal value and storing the corresponding index, much in the same way the Viterbi algorithm is a modification to the Forward-Backward algorithm.

4.6.1 Reduced search space

The algorithm is, however, intractable in this case, due to the extremely large size of the solution space. This is dealt with by reducing the search space: only a small number of most likely notes for every analysis frame are taken into account. First, at most K most salient pitches in every frame are selected if their salience is higher than the threshold calculated as the crossing point of the active-note and the inactive-note salience models (in our case 0.03); then, every possible k-combination of the selected notes if created, where $k = 1, \ldots, K$ and evaluated with the salience model; finally, L note combinations with the highest likelihood according to the salience model are selected and used in the frontier decoder.

To reduce the effect of short-time salience fluctuations, the salience matrix was smoothed before selecting the most salient pitches, by applying a single-pole IIR filter to each pitch with the same parameter a. The most optimal value of a was determined experimentally and set to 0.5 (see Fig. 12).

4.6.2 Acoustic model

To obtain the note saliences, we have used the harmonic NMF model proposed in [22] as the acoustic model, with tempo-synchronous analysis frame size of $1/6^{\text{th}}$ of a beat.

4.6.3 Fudging

The salience model was used with an exponential fudge factor κ, balancing the influence between the musicological and the salience model. Additionally, the full musicological model $\mathrm{P}_F(\mathbf{N})$ was mixed with the simple pitch-dependent Bernoulli model $\mathrm{P}_B(\mathbf{N})$ from subsection 4.5.2, with a complementarily weighting factor μ.

$$\widehat{\mathbf{N}} = \arg\max_{\mathbf{N}} \mathrm{P}(\mathbf{S}|\mathbf{N})^{\kappa} \mathrm{P}_B(\mathbf{N})^{\mu} \mathrm{P}_F(\mathbf{N})^{(1-\mu)}. \tag{36}$$

4.6.4 Evaluation metric

All multiple pitch estimation results were evaluated using onset-based $\mathcal{F}$-measure, similarly to the evaluation used in the MIR Exchange (MIREX), an annual MIR algorithm evaluation campaign [3]. The $\mathcal{F}$-measure is calculated as a harmonic

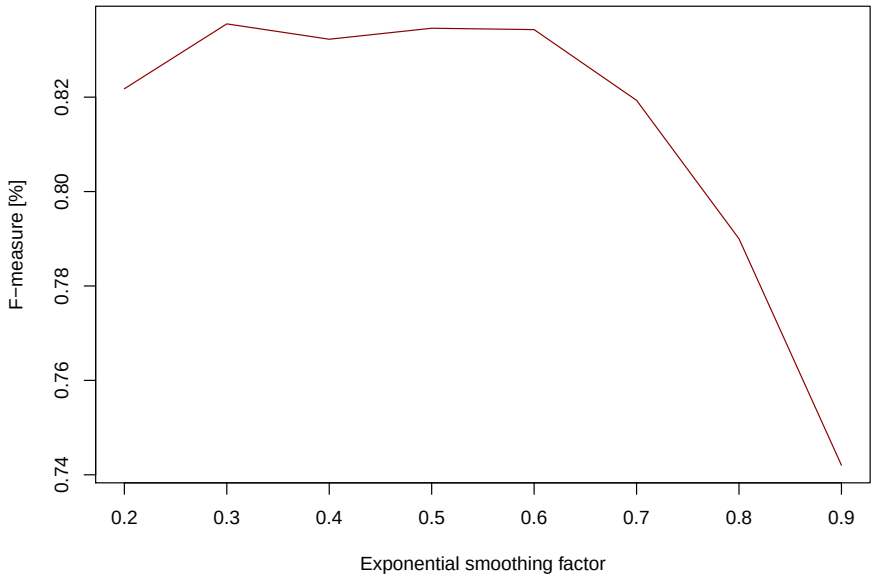


Figure 12: Onset-based $\mathcal{F}$-measure against the exponential smoothing factor a.

B	D	N	P	V	HC	HCV	HCDVPN
79.3%	???	79.7%	79.6%	82.3%	79.9%	82.3%	82.2%

Table 3: Maximal $\mathcal{F}$-measure values obtained for the tested models, compared with the baseline Bernoulli (B) model.

mean between the precision $\mathcal{P}$ (ratio of the number of correctly detected notes to all detected notes) and the recall $\mathcal{R}$ (ratio of the number of correctly detected notes to the number of ground-truth notes). A note was considered correctly detected if the pitch was exactly correct and the detected onset was within 6 frames (1 beat) from the correct onset position. The detected offset was ignored, as it is generally believed that an accurate estimation of the note offset time is in many cases extremely difficult if not impossible.

4.6.5 Results

Due to high computational demands of the proposed algorithm, all audio recordings were cut to 320 frames ($53^1/_3$ beats), which corresponds to roughly 30 seconds with the average frame length in our dataset of 93 ms (average tempo of about 108 beats per minute). The testing dataset was also reduced to 20 audio files.

The K and L parameters were chosen experimentally to have the values of 6 and 64 (2^6), respectively. The $\mathcal{F}$-measure was obtained for 6 different musicological models: HC, D, V, P, N, HCV and the full model: HCVDPN (actually HCVD, because the interpolation weights for P and N submodels were zero), for the values of the acoustic fudge factor κ between 0.5 and 2, and the values of the Bernoulli model weight between 0 and 1. The results are presented in Figs 13, 14 and 15 and summarized in Table 3.

The optimal value of κ is around 1, which suggests that the models were well trained. The full model, HCDVPN, i.e., interpolation of all proposed submodels (though with zero weights for two of them), has outpeformed the baseline pitch-dependend Bernoulli model (equivalent to thresholding with pitch-dependent threshold value) by 3%, but it is clear that the most important contributors to that improvement are the vertical models: V and D.

5 Conclusion

In this paper, we have proposed a probabilistic polyphonic pitch model that can be used for multiple pitch estimation. The model is a 3-layer DBN with 2 hidden layers corresponding to the chords and the notes. Modeling notes is done efficiently by means of linear and log-linear interpolation between simpler submodels, each responsible for modeling a different aspect of music.

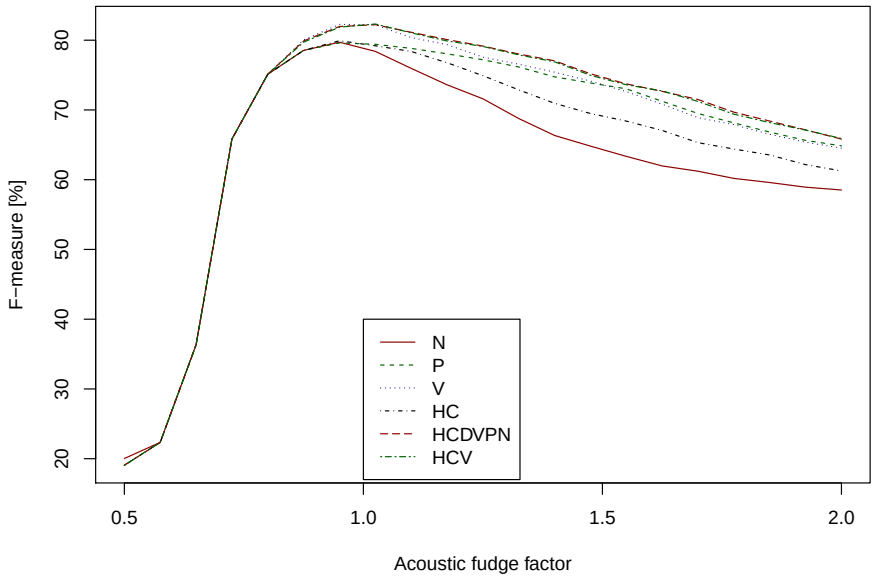


Figure 13: Onset-based $\mathcal{F}$-measure against the acoustic fudge factor κ for different musicological models. The optimal, different for each data point, value of μ was used.

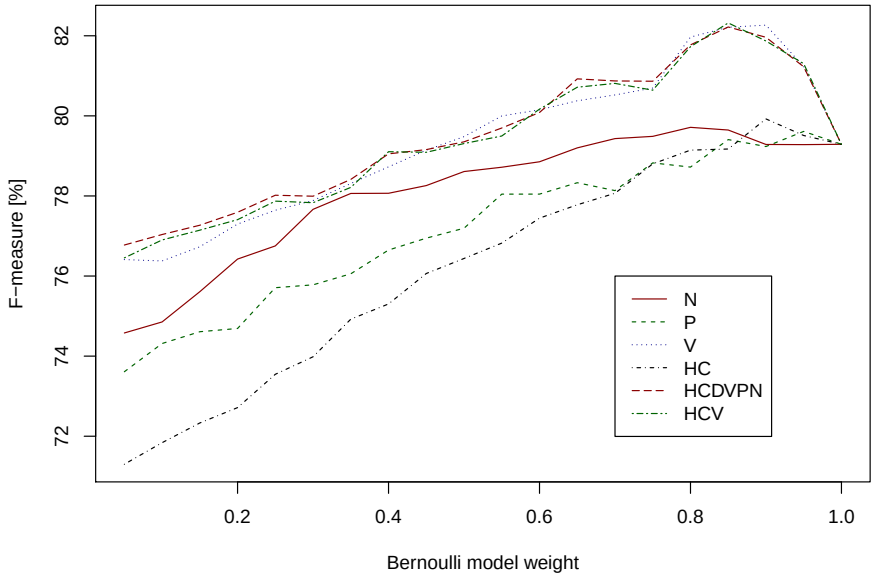


Figure 14: Onset-based $\mathcal{F}$-measure against the Bernoulli model weight μ. The optimal, different for each data point, value of κ was used.

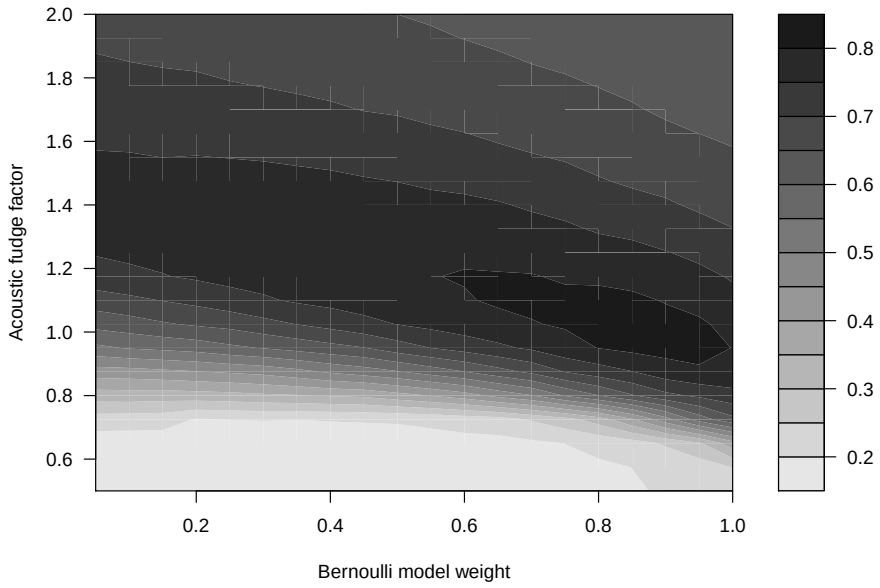


Figure 15: Contour plot of the $\mathcal{F}$-measure obtained for different values of μ and κ for the HCDVPN model.

The proposed framework was evaluated two-fold: in the symbolic experiments we have observed the modeling power represented in values of the cross-entropy and the para-cross-entropy; in the acoustic experiments we have performed actual multiple pitch estimation with our proposed model, using a harmonic NMF model as the acoustic model. In both experiments the proposed model offered an improvement over the baseline technique, i.e., a Bernoulli model (equivalent to thresholding of the salience). Analyzing the cross-entropies also showed that it is beneficial to combine submodels by means of interpolation, as adding models decreases the cross-entropy, especially the para-cross-entropy for notes. Using log-linear interpolation, although computationally more demanding (due to the need of re-normalization of the composite model), offered higher performance than the linear interpolation.

6 Acknowledgments

This work is supported by INRIA under the Associate Team Program VERSAMUS (`http://versamus.inria.fr/`).

References

[1] S.A. Abdallah and M.D. Plumbley. Unsupervised analysis of polyphonic music by sparse coding. *Neural Networks, IEEE Transactions on*, 17(1):179–196, 2006.

[2] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Discriminative non-negative matrix factorization for multiple pitch estimation. In *Proc. 13th International Conference on Music Information Retrieval (ISMIR)*, 2012.

[3] J.S. Downie. The music information retrieval evaluation exchange (MIREX). *D-Lib Magazine*, 12(12):795–825, 2006.

[4] C. Févotte, N. Bertin, and J.L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: with application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[5] S. Flossmann, M Grachten, and G. Widmer. Expressive performance rendering with probabilistic models. In A. Kirke and E.R. Miranda, editors, *Guide to Computing for Expressive Music Performance*, pages 75–98. Springer London, 2013.

[6] S. Fukayama, K. Nakatsuma, S. Sako, T. Nishimoto, and S. Sagayama. Automatic song composition from the lyrics exploiting prosody of the Japanese language. In *Proc. 7th Sound and Music Computing Conference (SMC)*, pages 299–302, 2010.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. In *Proc. 4th International Conference on Music Information Retrieval (ISMIR)*, pages 229–230, 2003.

[8] F. Jelinek and R.L. Mercer. Interpolated estimation of markov source parameters from sparse data. *Pattern recognition in practice*, pages 381–397, 1980.

[9] H. Kameoka, T. Nishimoto, and S. Sagayama. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Trans. Audio, Speech, and Language Processing*, 15(3):982–994, 2007.

[10] H. Kaneko, D. Kawakami, and S. Sagayama. Functional harmony annotation database for statistical music analysis. In *Demonstration at 11th International Conference on Music Information Retrieval (ISMIR)*, 2010.

[11] T.H. Kim, S. Fukayama, T. Nishimoto, and S. Sagayama. Performance rendering for polyphonic piano music with a combination of probabilistic models for melody and harmony. In *Proc. of Sound and Music Computing Conference (SMC)*, pages 23–30, 2010.

[12] D. Klakow. Log-linear interpolation of language models. In *Proc. 5th International Conference on Spoken Language Processing*, pages 1695–1699, 1998.

[13] A. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. 7th International Conference on Music Information Retrieval (ISMIR)*, pages 216–221, 2006.

[14] A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Trans. Audio, Speech, and Language Processing*, 16(2):255–266, 2008.

[15] K.P. Murphy. *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, University of California, 2002.

[16] B. Niedermayer. Non-negative matrix division for the automatic transcription of polyphonic music. In *ISMIR 2008, 9th International Conference on Music Information Retrieval*, pages 544–545, 2008.

[17] H. Papadopoulos and G. Peeters. Large-scale study of chord estimation algorithms based on chroma representation and HMM. In *Proc. International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 53–60. IEEE, 2007.

[18] J. Paulus and A. Klapuri. Music structure analysis using a probabilistic fitness measure and an integrated musicological model. In *Proc. 9th International Conference on Music Information Retrieval (ISMIR)*, pages 369–374, 2008.

[19] J. Pauwels and J. Martens. Integrating musicological knowledge into a probabilistic framework for chord and key extraction. In *In proc. 128th AES Convention*, page 9, 2010.

[20] The Mutopia Project. Free classical and contemporary sheet music. `http://www.mutopiaproject.org/`, March 2011.

[21] S.A. Raczyński, N. Ono, and S. Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proc. 8th International Conference Music Information Retrieval (ISMIR)*, pages 381–386, 2007.

[22] S.A. Raczyński, N. Ono, and S. Sagayama. Extending nonnegative matrix factorization—a discussion in the context of multiple frequency estimation of musical signals. In *Proc. 17th European Signal Processing Conference (EUSIPCO)*, pages 934–938, 2009.

[23] S.A. Raczynski, E. Vincent, F. Bimbot, and S. Sagayama. Multiple pitch transcription using dbn-based musicological models. In *Proc. 11th International Conference on Music Information Retrieval (ISMIR)*, pages 363–368, 2010.

[24] C. Raphael and J. Stoddard. Harmonic analysis with probabilistic graphical models. In *Proc. 4th International Conference on Music Information Retrieval (ISMIR)*, pages 177–181, 2003.

[25] M.P. Ryynanen and A. Klapuri. Polyphonic music transcription using note event modeling. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 319–322. IEEE, 2005.

[26] M.P. Ryynänen and A.P. Klapuri. Modelling of note events for singing transcription. In *Proc. ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*, 2004.

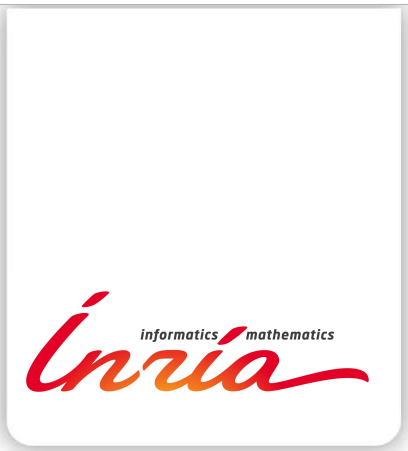
[27] A. Shirai and T. Taniguchi. A proposal of an interactive music composition system using gibbs sampler. *Human-Computer Interaction. Design and Development Approaches*, pages 490–497, 2011.

[28] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, 2003.

[29] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5518–5521. IEEE, 2010.

[30] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech and Language Processing*, 18(3):528–537, 2010.

[31] E. Vincent, S.A. Raczyński, N. Ono, and S. Sagayama. A roadmap towards versatile MIR. In *Proc. f(MIR) session of 11th International Conference on Music Information Retrieval (ISMIR)*, pages 662–664, 2010.



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-0803