Dynamic Bayesian networks for symbolic polyhonic pitch modeling

Stanisław Andrzej Raczyński,[†1]
Emmanuel Vincent[†2] and Shigeki Sagayama[†1]

The performance of many MIR analysis algorithms, most importantly polyphonic pitch transcription, can be improved by introducing musicological knowledge to the estimation process. We have developed a probabilistically rigorous musicological model that takes into account dependencies between consequent musical notes and consequent chords, as well as the dependencies between chords, notes and the observed note saliences. We investigate its modeling potential by measuring and comparing the cross-entropy with symbolic (MIDI) data.

1. Introduction

Symbolic pitch modeling, i.e. modeling the prior distribution of note sequences $P(\mathbf{N})$, also known as musicological modeling, is the equivalent of language modeling in speech processing. It has the potential to be used many in Music Information Retrieval (MIR) tasks, like multiple pitch estimation, algorithmic composition, computational musicology, symbolic music analysis, music segmentation, etc., as a part of an integrated statistical model of music[1)].

For example, in the task of polyphonic music transcription, i.e. estimating the pitches, the onset times and the durations of the musical notes present in a recorded audio signal, incorporating a symbolic pitch model would mean a transition from a ML-like estimation

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

to estimating the notes in the MAP sense:

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

where $\mathrm{P}(\mathbf{S}|\mathbf{N})$ is an acoustic model, such as Nonnegative Matrix Factorization[2),3),4)], Matching Pursuit, or other model[5),6),7),8),9),?)]. While acoustic modeling has been widely studied, symbolic 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[10)] proposed a melody transcription method that uses a Hidden Markov Model (HMM) together with a simple musical key model. Their approach however is limited in the sense that it models only monophonic note sequences. Because of that, their approach lacks modeling of the dependencies between concurrent pitches. Raphael and Stoddard[11)] proposed to use an HMM as a symbolic model for harmonic analysis, i.e. estimating 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.[12)]). These HMM-based approaches, however, lack absolute pitch modeling and the temporal dependencies are only present between chords.

In this paper we propose a single probabilistic pitch model based on Dynamic Bayesian Networks (DBNs). We model both the dependencies between consequent notes (harmony) and the temporal dependencies between notes and chords. The prior over the note activities $\mathrm{P}(\mathbf{N})$ models the temporal dependencies between the hidden variables (similar to those of an HMM) and includes a hidden layer of variables representing chords.

2. Model

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

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

The corresponding network structure is presented in Fig. 1.

However, 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 train or be used for inference in practice, as it requires $2^{88} \times 2^{88} \times 24$ (assuming the full piano keyboard) distinct probability values to be fully defined. To deal with this

†1 The University of Tokyo
†2 Institut National de Recherche en Informatique et en Automatique

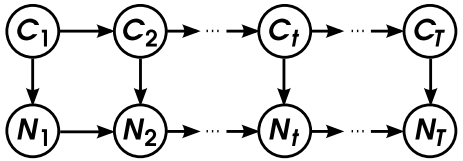

Fig. 1 Proposed structure of the Dynamic Bayesian Network for polyphonic pitch modeling.

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

$$\mathrm{P}(\mathbf{N}_t|\mathbf{N}_{t-1}, C_t) = \prod_{k=1}^{K} \mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}, C_t, N_{t,1}, N_{t,2}, \ldots, N_{t,k-1}), \tag{4}$$

where k is the analyzed pitch and $K = 88$ is the size of the analyzed pitch range. This procedure helps to reduce the dimensionality of the note combination variable, but the resulting formula is still difficult to apply in practice due to the highly dimensional conditioning variable set. We therefore split this distribution into three separate ones by using the following approximation:

$$\begin{aligned}&\mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}, C_t, N_{t,1}, N_{t,2}, \ldots, N_{t,k-1}) \qquad (5)\\ &\quad \approx Z^{-1} \cdot \mathrm{P}(N_{t,k}|\mathbf{N}_{t-1})^{\lambda_1} \cdot \mathrm{P}(N_{t,k}|C_t)^{\lambda_2} \mathrm{P}(N_{t,k}|N_{t,1}, N_{t,2}, \ldots, N_{t,k-1})^{\lambda_3},\end{aligned}$$

where $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ are exponential weights, and Z is the normalization factor:

$$Z = \sum_{N_{t,k}=0}^{1} \mathrm{P}(N_{t,k}|\mathbf{N}_{t-1})^{\lambda_1} \mathrm{P}(N_{t,k}|C_t)^{\lambda_2} \cdot \tag{6}$$

$$\cdot\, \mathrm{P}(N_{t,k}|N_{t,1}, N_{t,2}, \ldots, N_{t,k-1})^{\lambda_3}. \tag{7}$$

This kind of log-linear interpolation of musicological models was first proposed by Klakow in the context of language modeling[13]. In this paper, each of the separated musicological models is called a *submodel* and each of them is responsible for modeling a different musicological aspect of the note sequences: voice movement and note duration, harmony and polyphony, respectively.

The exponential weights control the influence of each of the submodels on the resulting note activity distribution: setting them to zero effectively disables the corresponding submodel by making it completely uniform, while using higher values means that the corresponding models will have bigger impact on the results.

3. Training

The parameters of all submodels were trained by counting occurrences. In all cases a simple smoothing procedure was used: every event was assumed to have occurred at least once.

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[12]. 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 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 model

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

$$\mathrm{P}(N_{t,k}|C_t) = \mathrm{P}(\,\mathrm{inter}\{k; \mathrm{root}\{C_t\}\}|\mathrm{mode}\{C_t\}), \tag{8}$$

where inter is the musical interval operator, root is the root note operator and

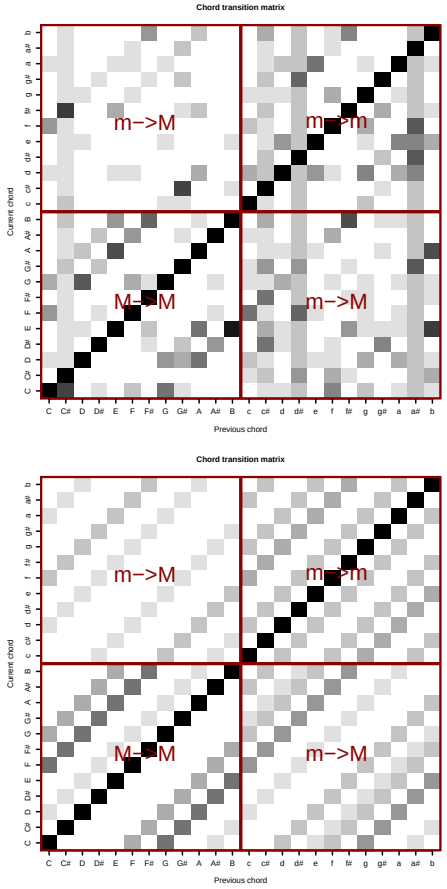

Fig. 2 Chord transition probability matrix if state tying was not used (left) and if transition probabilities were tied (right). Darker color represent higher probability values. Minor chords (m) are annotated with lower case and major chords (M) with upper case.

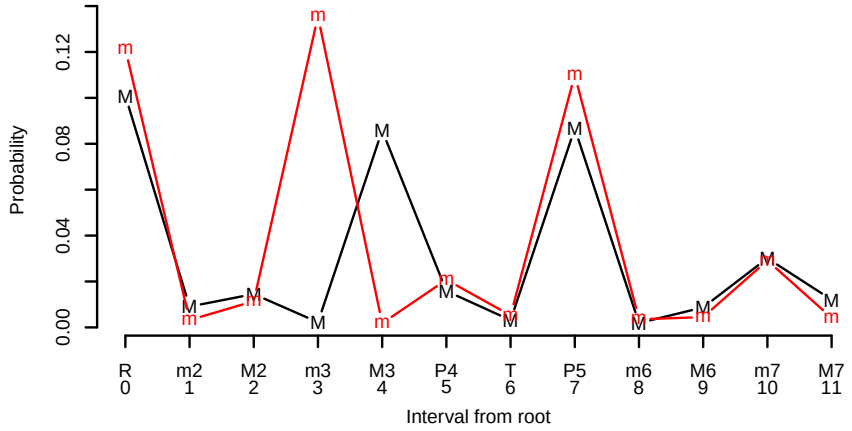

Fig. 3 Pitch probability distribution for major (M) and minor (m) chords as a function of the interval from the chord's root note.

mode is the mode operator, i.e. major or minor. The corresponding probability distribution is presented in Fig. 3.

3.3 Voice model

The voice model $\mathrm{P}(N_{t,k}|\mathbf{N}_{t-1})$ also suffers from the high dimensionality of the conditioning variable set. To reduce the dimensionality, we have proposed three different, simplified models that assume that only some pitches are relevant in modeling the voice movement.

3.3.1 Reduced voice model

In this model we assume dependence only on pitches from a small range $(-R,\ldots,0,1,\ldots,R)$, relative to the current pitch k:

$$\mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}) \approx \mathrm{P}(N_{t,k}|N_{t-1,k-R}, N_{t-1,k-R+1},\ldots,N_{t-1,k},\ldots,N_{t-1,k+R}), \quad (9)$$

This is based on an assumption that voice movement is limited to small jumps, typically within a single octave, i.e. $R = 12$.

3.3.2 Duration-only model

In this model individual note activities are assumed to be dependent only on

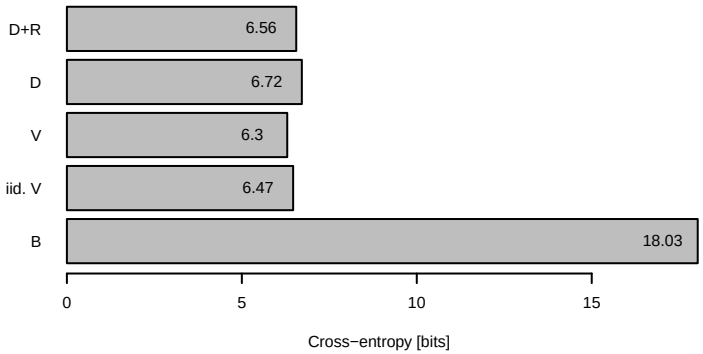

Fig. 4 Cross-entropy obtained for different voice models and an unconditional Bernoulli model (B) for reference.

the previous state of the same pitch:

$$\mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}) \approx \mathrm{P}(N_{t,k}|N_{t-1,k}), \tag{10}$$

i.e. by a pitch-dependent conditional Bernoulli model.

3.3.3 Voice movement model

In this model we use the following approximation:

$$\mathrm{P}(N_{t,k}|\mathbf{N}_{t-1}) \approx \mathrm{P}(N_{t,k}|M_{t,k}), \tag{11}$$

where $M_{t,k}$ is the distance between the pitch k and the closest active pitch in the previous time frame. 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.

3.3.4 Comparison

The models are compared by the cross-entropies of the observed note data $\mathbf{N}$ (see section 5), calculated on the testing data set (see section 4). We have compared the duration-only model (D), the reduced voice (R) and the duration-only model models combined by log-linear interpolation (R+D), and two versions of the voice movement model: independent voice movement model (V) and an independent version identically distributed for all pitches (iid. V). For a baseline we have juxtaposed these models with an unconditional Bernoulli model (B). The results are presented in Fig. 4.

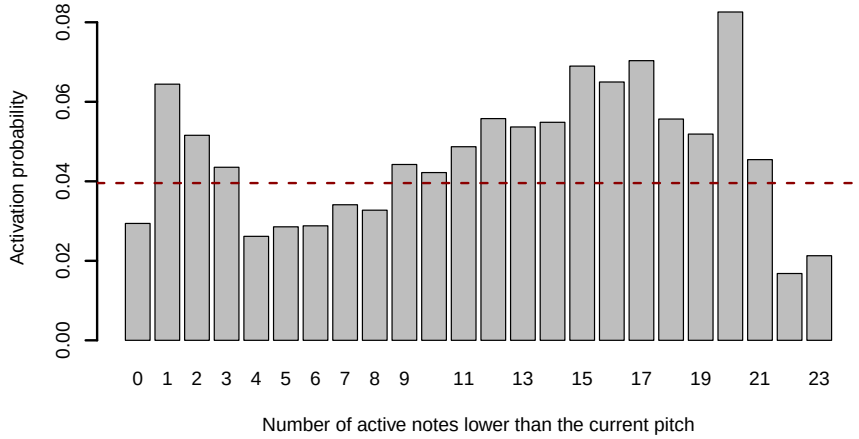

Fig. 5 Polyphony model $\mathrm{P}(N_{t,k}|L_{t,k})$. The dashed line marks the marginal note activity probability $\mathrm{P}(N_{t,k})$.

3.4 Polyphony model

The probability distribution $\mathrm{P}(N_{t,k}|N_{t,1}, N_{t,2}, \ldots, N_{t,k-1})$ is more difficult to model due to high dimensionality of the conditioning variable set. However, since global note distributions and pitch are already modeled by the duration and the harmony model, respectively, the polyphony model can be simplified to model only the number of notes active simultaneously:

$$\mathrm{P}(N_{t,k}|N_{t,1}, N_{t,2}, \ldots, N_{t,k-1}) \approx \mathrm{P}(N_{t,k}|L_{t,k}), \tag{12}$$

where $L_{t,k} = \sum_{m=1}^{k-1} N_{t,m}$. The resulting distribution is plotted in Fig. 5.

3.5 Exponential weights

The interpolation weights λ from Eq. 5 are optimized by maximizing their log-likelihood:

$$\widehat{\lambda} = \arg\max_{\lambda} \log \mathrm{P}(\mathbf{N}|\lambda), \tag{13}$$

which is calculated for the validation dataset (see section 4). Optimization is performed using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method (a quasi-Newton optimization), built in the GNU R environment as the `optim()` function[14]. The resulting values are listed in Table 1. The initial coefficient

Coefficient	Model	H	HC	P	V	VP	HVP	HCVP
λ_1	Harmony	0.969	0.979	—	—	—	0.014	0.024
λ_2	Voice	—	—	0.966	—	0.954	0.951	0.951
λ_3	Polyphony	—	—	—	1.016	0.028	0.019	0.022

Table 1 Trained values of exponential coefficients for submodels from Eq. 5.

values were all equal to $\lambda_p = 1$, $p \in \{1, 2, 3\}$.

4. Data

Two datasets were used in the experiments: the widely used RWC database[15] and the Mutopia Project dataset[16]. 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[17]. 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 submodel weights from Eq. 5 were trained on the validation set. Experiments were performed on the testing data. All symbolic data was quantized and 1 frame corresponded to $1/6^{\text{th}}$ of a beat, so it was rather sheet music encoded in MIDI format than performance data, as is the case with the original RWC data.

5. Symbolic evaluation

The models are compared by calculating the cross-entropy of the observed note data given the musicological prior. We compute the marginal cross-entropies, i.e. cross-entropies with the chord sequence marginalized out:

$$\begin{aligned} \mathrm{H}(\mathbf{N}) &= -\frac{1}{T}\log_2 \mathrm{P}(\mathbf{N}|\lambda) \\ &= -\frac{1}{T}\log_2 \sum_{\mathbf{C}_O} \mathrm{P}(\mathbf{N}|\mathbf{C}_O, \lambda)\mathrm{P}(\mathbf{C}_O|\lambda). \end{aligned} \tag{14}$$

The conditional cross-entropy $\mathrm{H}(\mathbf{N}|\mathbf{C})$ is calculated by summing over all possible chord sequences. This is calculated using the frontier algorithm[18], which is the DBN equivalent of the forward-backward algorithm. Three reference models were used for comparison: the uniform model $\mathrm{P}(N_{t,k} = 1) = 1/2$, the independent and identically distributed model $\mathrm{P}(N_{t,k}) \sim \text{Bernoulli}(p)$ with $p = 0.03986$ and an independent model $\mathrm{P}(N_{t,k}) \sim \text{Bernoulli}(p_k)$. The values of p_k are shown in Fig. 6.

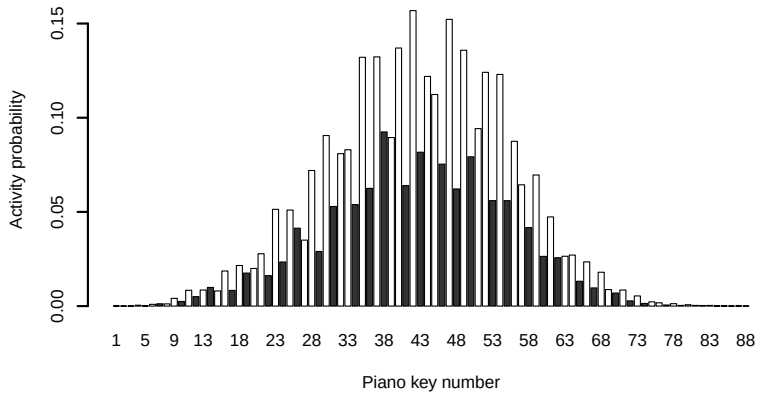

Fig. 6 Parameters p_k of the independent note activity model. Black and white bars correspond to black and white piano keys, respectively.

The results are presented in Fig. 7. They how that modeling harmonic relations between pitches (HC) leads to better modeling performance than modeling prior pitch distributions with a Bernoulli model (i.). However, modeling horizontal relations between pitches (V) seems more to be important than the vertical harmony models. Combining all proposed model (HCVP) results in the lowest note entropy.

6. Conclusion and future work

We have presented a probabilistic symbolic pitch model that is a log-linear interpolation of a number of simpler models representing complementary properties of a note sequence. The proposed models have been evaluated by calculating the marginal cross-entropy $\mathrm{H}(\mathbf{N})$ of the testing data set given the model. The entropy of the observed notes is as low as 6.44 bits per time frame, compared to the reference unconditional Bernoulli model (18.03 bits) and the uniform model (88 bits).

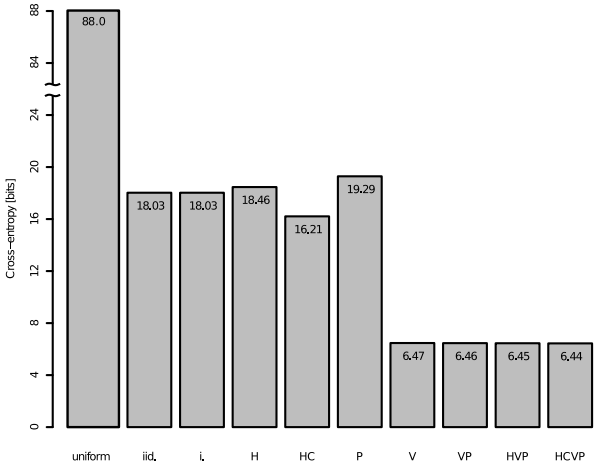

Fig. 7 Average marginal cross-entropies for the tested models. iid. = iid. Bernoulli, i. = independent Bernoulli, H = Harmony model, C = chord model, P = Polyphony model, V = Voice model.

In future work we will focus on combining the proposed models with other MIR models from[1], e.g. an acoustic model to perform multiple pitch estimation, so as to observe the practical implications of incorporating a musicological prior to the estimation process.

7. Acknowledgments

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

References

1) E. Vincent, S.A. Raczyński, N. Ono, and S. Sagayama. f(MIR): a roadmap towards versatile MIR. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 662–664, 2010.
2) S.Raczyński, N.Ono, and S.Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proc. International Conference Music Information Retrieval (ISMIR)*, pages 381–386, 2007.
3) P.Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180, 2003.
4) 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.
5) 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.
6) M.P. Ryynänen and A.Klapuri. Polyphonic music transcription using note event modeling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 319–322, 2005.
7) S.A. Abdallah and M.D. Plumbley. Unsupervised analysis of polyphonic music by sparse coding. *IEEE Trans. Neural Networks*, 17(1):179–196, 2006.
8) A.Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 216–221, 2006.
9) 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.
10) M.P. Ryynänen and A.P. Klapuri. Modelling of note events for singing transcription. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2004.
11) C.Raphael and J.Stoddard. Harmonic analysis with probabilistic graphical models. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 177–181, 2003.
12) Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5518–5521, 2010.
13) D.Klakow. Log-linear interpolation of language models. In *Fifth International Conference on Spoken Language Processing*, volume5, pages 1695–1698, 1998.
14) R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.
15) M.Goto, H.Hashiguchi, T.Nishimura, and R.Oka. RWC music database: Music genre database and musical instrument sound database. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 229–230, 2003.
16) The Mutopia Project. Free classical and contemporary sheet music. `http://www.mutopiaproject.org/`, March 2011.
17) H. Kaneko, D. Kawakami, and S. Sagayama. Functional harmony annotation database for statistical music analysis. In *Demonstration at International Conference on Music Information Retrieval (ISMIR)*, 2010.
18) K.P. Murphy. *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, University of California, 2002.