Automatic Transcription of Bell Chiming Recordings

Matija Marolt, *Member, IEEE*

***Abstract*— Bell chiming is a folk music tradition that involves performers playing rhythmic patterns on church bells. The paper presents a method for automatic transcription of bell chiming recordings, where the goal is to detect the bells that were played and their onset times. We first present an algorithm that estimates the number of bells in a recording and their approximate spectra. The algorithm uses a modified version of the intelligent k-means algorithm, as well as some prior knowledge of church bell acoustics to find clusters of partials with synchronous onsets in the time-frequency representation of a recording. Cluster centers are used to initialize non-negative matrix factorization that factorizes the time-frequency representation into a set of basis vectors (bell spectra) and their activations. To transcribe a recording, we propose a probabilistic framework that integrates factorization and onset detection data with prior knowledge of bell chiming performance rules. Both parts of the algorithm are evaluated on a set of bell chiming field recordings.**

***Index Terms*— Audio systems, Bell chiming, Music transcription, Signal analysis**

I. Introduction

BELL chiming is a folk music tradition that still exists in its original context today. It takes place in the church tower and its original role is strongly connected to Christian religious contexts. Bell chiming combines the signaling, ritual, and musical functions, because it is most often used to call the faithful to mass in a musical way, and at the same time to mark important church holidays. This is how the difference between conventional bell ringing and bell chiming as a more solemn form of playing the bells is established [1].

Slovenian-style bell chiming is performed by musicians holding the clapper and striking the rim of the stationary bell at regular intervals. The sound is thus not produced by a swinging bell hitting the clapper, but by the clapper, typically held close to the rim, hitting the bell's rim. This gives musicians more control in altering the rhythm, speed, dynamics and accents of individual strikes, as well as leaving out strikes if desired. In the so called "Flying" tunes, one of the bells (usually the largest) is swung with a rope or electronically, and all the other bells, which are stationary, are played by striking the clapper. As a rule, each musician is responsible for playing one bell, and should strike the bell only with its clapper (touching the bell's rim with hands or other tools is not allowed). Another important rule in bell chiming is that two tones can never be played at the same time (but exceptions do occur).

Bell chiming tunes contrast one another in the method of playing, the number of bells used, and their rhythmic and metric structure. Tunes themselves consist of repeated rhythmic patterns into which various changes, typically dynamic and agogic are included to enliven the performance. Since musicians perform in groups, without the group's consent, only small changes are possible within the time limits allocated to the bell chimer for performing his role. These changes are usually expressed as double strikes, triplets, or pauses [1]. Pioneering work in analysis of Slovenian bell chiming practices was made by Ivan Mercina in the late 19th and early 20th century, who introduced a numerical notation system and published a repertoire of 243 bell chiming tunes. His work is carried on by researchers of the Institute of Ethnomusicology of the Scientific Research Centre of the Slovenian Academy of Sciences and Arts, who are still actively researching bell chiming practices. Their digital archive of Slovenian folk music and dances EthnoMuse [2] holds a large collection of bell chiming recordings, collected from the 1950s onwards. Only parts of the archive have been manually transcribed and annotated.

In this paper, we present a method for automatic transcription of bell chiming recordings. Automatic music transcription is a difficult problem, although methods are improving constantly; Klapuri and Davy provide an extensive overview of the field [3]. Lately, non-negative matrix factorization (NMF) and related techniques, such as probabilistic latent component analysis, are becoming an increasingly popular choice for music information retrieval tasks such as sound source separation [4], alignment of audio to score [5], as well as multi-pitch estimation and music transcription [6-12]. For transcription, a time-frequency representation of the audio signal is decomposed into a product of basis vectors (spectral templates) and their time-varying activations. Basis vectors can be interpreted as note spectra, while activations give information on onsets and offsets of notes. To improve factorization, most authors use

Manuscript received 30.9.2010. This work was supported in part by the Slovenian Government-Founded R&D project EthnoCatalogue: creating semantic descriptions of Slovene folk song and music.

Matija Marolt is with University of Ljubljana, Faculty of Computer and Information Science, Trzaska 25, 1000 Ljubljana (phone: +386 1 4768483; fax: +386 1 4264647; e-mail: matija.marolt@ fri.uni-lj.si).

prior knowledge to put constraints on the learning process, including limiting the basis vectors to a set of predefined or learned templates [10-12], sparsity [7], harmonicity [8, 13] and temporal smoothness [8].

In this paper we propose two novel techniques that, combined with non-negative matrix factorization, can be used to transcribe bell chiming recordings. In section II we describe an algorithm for estimating the number of bells and their approximate spectra in a recording by using prior knowledge of church bell acoustics and bell chiming performance rules. We show that these estimates can be used to initialize NMF, leading to a meaningful factorization, suitable for transcription. In section III, we propose a transcription algorithm, which is based on a Markov model integrating factorization and onset detection data with prior knowledge of bell chiming performance rules. Both proposed algorithms are evaluated in section IV. Section V concludes the paper and gives ideas for future work.

II. Decomposition

When non-negative matrix factorization (NMF) is applied to transcription of polyphonic music, a time-frequency representation **V** (of size *F*x*T*) is approximated by a product of two non-negative matrices **W** and **H**, so that:

$$\mathbf{V} \cong \mathbf{WH}\,, \tag{1}$$

where **W** is a matrix of basis vectors (of size *F*x*M*) and **H** a matrix of coefficients (of size *M*x*T*). In music transcription, the columns of **W** correspond most naturally to individual music events (their spectra), while the rows of **H** explain how activations (amplitudes) of these events change over time.

The naive approach of applying NMF to transcription of music signals simply by factorizing the magnitude or power spectrum has several shortcomings. As music events overlap in time, there is no guarantee that NMF will separate individual music events into separate basis vectors. A single vector may end up containing components of several events or only a subset of components of an individual event. We also have to set the number of basis vectors in advance; using too few basis vectors results in vectors containing several events, on the other hand, too many vectors may result in fragmentation of events over several vectors.

It is therefore necessary to constrain the NMF learning process to yield meaningful factorizations – meaningful denoting the fact that basis vectors should represent the spectra of individual events in a recording. Usually, prior knowledge of instruments played, either in the form of instrument models or harmonicity constraints, is used to constrain the learning process. For piano transcription, Niedermayer [11] used a preset number of fixed basis vectors learned from recordings of individual piano notes and only adapted the matrix of coefficients **H** during learning. Raczynski et al. [10] used a preset number of basis vectors corresponding to individual piano notes, initialized and constrained the vectors to non-zero values only for frequency bins corresponding to perfectly harmonic partial series and learned the weights of these harmonics, as well as the matrix of coefficients. Vincent et al. [9] and Bertin et al. [8] imposed harmonicity constraints on the basis vectors, while Grindlay and Ellis [12] constrained the vectors to a model space previously learned on a set of instruments.

Church bell sounds are complex and inharmonic, bells are cast and tuned individually for each church, so little can be assumed of the sound of bells in a given recording in advance. It is therefore difficult to constrain the learning process, because we cannot learn a complete set of bell spectral templates in advance (as in [10-12]) or impose other constraints such as harmonicity [8, 9]. However, we show that we can obtain meaningful factorizations of church bell recordings, if NMF is properly initialized. To this end, we propose an algorithm for extraction of the number of bells and their approximate spectral templates from a recording, whereby prior knowledge of church bell acoustics and bell chiming performance rules is utilized. The obtained templates are used to initialize NMF, resulting in meaningful factorizations. In this section, we first provide some background on church bell acoustics and then outline the proposed algorithm.

A. Church Bell Acoustics

The shape or profile of a bell determines the relative frequencies of its vibrations. The conventional western shape of bells, which stems from the middle ages, tends to give the bell a single dominant pitch. Fig. 1 shows the magnitude spectrum of a bell, whose dominant pitch lies at 412 Hz. The names of several significant partials of the bell are shown, although (as can be seen) many others exist. The dominant pitch of the bell is defined by relations of three of its significant partials: *nominal*, *superquint* and *octave nominal* [14]. These form a near harmonic series with ratios 2/2, 3/2 and 4/2 resulting in a perceived virtual pitch at approximately half the nominal frequency. Most of the other partials, including the strongest for this particular bell (tierce), do not belong to this harmonic series.

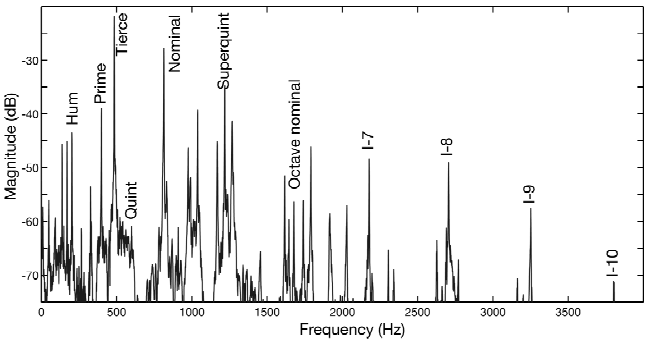

Fig. 1. Magnitude spectrum of a bell.

An extensive analysis of acoustics of church bells of western shape was made by Hibbert [14]. He showed that a linear relationship exists between positions of significant partials above the nominal (superquint, I-7, I-8 etc.) and the ratio of octave nominal to nominal frequency. Hence, a linear regression model can be used to infer the frequencies of these partials, if frequencies of the nominal and octave nominal are

known [14]. Frequencies of partials below the nominal do not exhibit such relationships, and their positions can only be coarsely approximated, as shown in TABLE I. The table shows positions of significant bell partials relative to the nominal; we calculated these positions on a set of 318 church bells.

TABLE I
POSITIONS AND MAGNITUDES OF PARTIALS RELATIVE TO THE NOMINAL

Partial	Mean and st. dev. of partial positions (cents)	Mean magnitude (dB)
hum	-2327 ± 84	-12
prime	-1247 ± 88	-7
tierce	-883 ± 41	-3
quint	-456 ± 83	-21
nominal	0 ± 0	0
superquint	684 ± 37	-2
oct. nominal	1221 ± 139	-4
I-7	1714 ± 87	-7
I-8	2095 ± 101	-9
I-9	2418 ± 115	-11
I-11	2933 ± 138	-17

Formally, if we know positions of the nominal and octave nominal partials and use Hibbert's regression model to estimate positions of partials above the nominal, we can define a bell spectral model containing the bell's significant partials as a Gaussian mixture:

$$M_{n,o}(x) = \sum_{k=1}^{4} a_k \, \mathrm{G}(x, n + \delta_k, \sigma_k) + \sum_{k=5}^{11} a_k \mathrm{G}(\mathrm{x}, p_{n,o}^k, \sigma_r) \ , \quad (2)$$

where n and o are frequencies of the nominal and octave nominal, a_k, δ_k and σ_k magnitudes, relative positions of bell partials and their standard deviations as given in TABLE I, $p_{n,o}^k$ the k-th partial position, σ_r the allowed deviation from the calculated position and G the unnormalized Gaussian function. $p_{n,o}^k$ is calculated with a linear regression model as:

$$p_{n,o}^k = (o - n) m_k + C_k \, , \quad (2)$$

where parameters m_k and C_k are taken from Table 5-4 of Hibbert's work [14, p. 105].

The time evolution of partials follows a roughly exponential decay curve, and is faster for higher partials, which on the other hand, are initially louder. Individual partials frequently exhibit beating, also evident in Fig. 2.

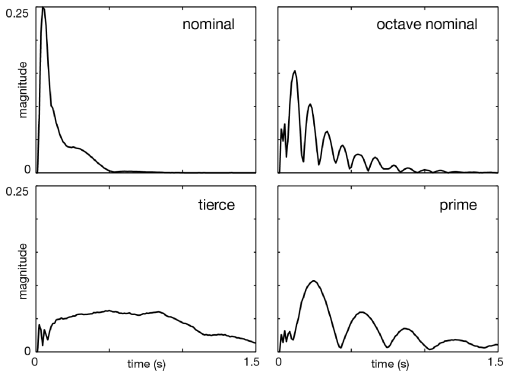

Fig. 2. Amplitude envelopes of four bell partials – beating is evident in the octave nominal and prime partials.

Beating is caused by the so-called doublets, which arise when bells are not symmetrical about a vertical axis through their centers. This asymmetry causes most of the vibrational modes in bells to split into two distinct modes with slightly different frequencies that beat against each other. Beating affects onset detection and decomposition of the time-frequency spectrogram, so we try to reduce its effects, as described in the following section.

B. Estimating the Number of Bells and Their Spectra

To estimate the number of bells and positions of their partials in a bell chiming recording, we consider the following. In a bell chiming performance, each bell is played many times, but bells are typically not played at the same time, as this is not allowed by the bell chiming performance rules. As a performance evolves, amplitude envelopes of partials of a bell exhibit similar behavior, especially at onsets – they are correlated. However, they are not correlated to amplitude envelopes of partials of other bells, because bells are not played simultaneously. We can make use of this *onset synchrony* of bell partials to find groups of partials of individual bells and thus find estimates of bell spectra.

The proposed algorithm is as follows. We first calculate the magnitude spectrogram **F** of a recording. To reduce variance in partial magnitudes in different frequency regions, we make use of the perceptual weighting model introduced by Vincent [15]. The algorithm multiplies the signal spectrum with weights that are inversely proportional to the specific loudness of the signal in individual gammatone filter modeling bands. However, instead of calculating and applying weights to each spectrogram frame, we calculate weights on the average signal spectrum (all frames are averaged) and multiply all spectrogram frames with the calculated weights. In this way, the overall shape of the spectrogram in time is not changed, and further steps of the algorithm, such as the calculation of the delta spectrogram, are not affected. Weighting results in a flattened time-frequency representation $\mathbf{F_w}$. Such flattening "amplifies" partials with small magnitudes, and ensures that the most energetic components do not dominate, thus enabling the bell finding algorithm to consider those partials when forming partial groups. Comparable methods for weighting the spectrum were also used by other authors. Klapuri [16] suggested to flatten the spectral energy distribution by scaling the bark scale sub-bands inversely proportional to their variance, while Virtanen [17] used a weighted cost function in which the observations were weighted so that the quantitative significance of the signal within each critical band was equal to its contribution to the total loudness.

As bells have sharp onsets and long decay times, the next step of the algorithm accentuates the fast positive changes (sharp onsets) in the magnitude spectrogram. The dynamics of changes within frequency bins of $\mathbf{F_w}$ are estimated by calculating the first order delta (regression) coefficients **D** of the bins with a sliding window of length N_d. Delta coefficients provide estimates of the gross shape of short time segments of

the spectrogram and are calculated with a least squares approximation as:

$$d_{ik} = (2\sum_{l=1}^{N_d} l^2)^{-1} \sum_{l=-N_d}^{N_d} l f_{ik+l}\ , \qquad (2)$$

where f_{ik} represents an element of $\mathbf{F_w}$ (i is the frequency index and k the time index). Half-wave rectification is used to prune negative values, and the resulting representation emphasizes fast and big changes, such as onsets, and deemphasizes slower and smaller changes, such as beating. This is illustrated in Fig. 3, which displays the amplitude envelope of a bell partial in a bell chiming recording (top) and its delta coefficients (bottom) – in the latter, onsets are emphasized and the effects of beating reduced. As we show in section IV, accuracy of the bell finding algorithm significantly increases when the delta spectrogram is used.

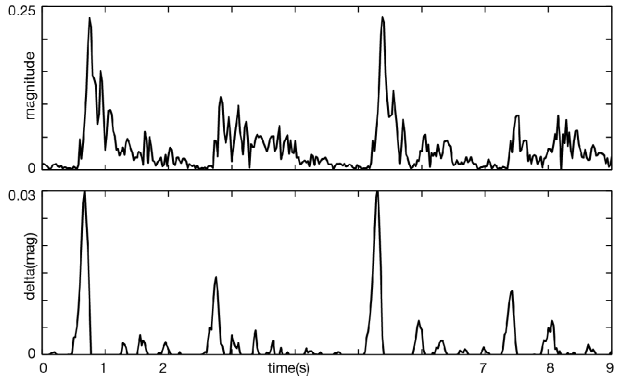

Fig. 3. Amplitude envelope and delta coefficients of a bell partial in a bell chiming performance

To discover groups of related partials, we first calculate covariances of their delta coefficients:

$$c_{ij} = \max\left(\frac{1}{n-1}\sum_{k=1}^{n}\left(d_{ik}-\mu_i\right)\left(d_{jk}-\mu_j\right), 0\right), \qquad (3)$$

where d_{ik} represents an element of the delta spectrogram $\mathbf{D}$ (i is the frequency index and k the time index) and μ_i the mean value of the i-th row of $\mathbf{D}$. Because delta coefficients emphasize onsets, the value c_{ij} represents a measure of *onset synchrony* of partials with frequencies corresponding to bins i and j. Partials of different bells do not overlap very often and bells are rarely struck at the same time, so a bell's partial has high onset synchrony with other partials of the same bell, but not with partials of other bells. A row $\mathbf{c}_i$ of the covariance matrix **C,** also called a *synchronicity pattern*, is therefore similar to the spectrum of the bell containing the partial corresponding to bin i.

If the covariance matrix is calculated globally on the entire spectrogram, bells which are dominant either due to their loudness, or the number of occurrences, are emphasized. Thus, less dominant bells may be underrepresented in **C** and missed by the grouping algorithm. This scenario can be avoided, if we calculate *local* covariance matrices $\mathbf{C}^{(t)}$, which stress locally important events. $\mathbf{C}^{(t)}$ are calculated on short segments of the delta spectrogram **D**, obtained with a sliding window of length N_c and a step size of $N_c/2$. They are merged into the *global* measure of onset synchrony **O** by weighting contributions of local synchronicity patterns with energies of partials in each segment, as approximated by $c_{ii}^{(t)}$:

$$o_{ij} = \frac{1}{\sum_t c_{ii}^{(t)}} \sum_t c_{ii}^{(t)} c_{ij}^{(t)} \quad . \qquad (4)$$

Fig. 4 displays two synchronicity patterns from the matrix **O**, calculated on a bell chiming recording with three bells. The top part of the figure shows onset synchrony of the tierce partial (488 Hz) of the bell from Fig. 1 (B1). Most of the bell's significant partials can be discerned and the obtained synchronicity pattern is similar to the bell's actual spectrum. Even though two more bells are present in the recording, amplitude envelopes of their partials are quite different and therefore not correlated with B1's tierce. The bottom part of the figure is placed at the nominal frequency of B1 (824 Hz), which coincides with the superquint of another bell with nominal frequency of 548 Hz (B2). Partials of both bells can be discerned, B1 being more dominant, as its nominal is stronger than the superquint of B2.

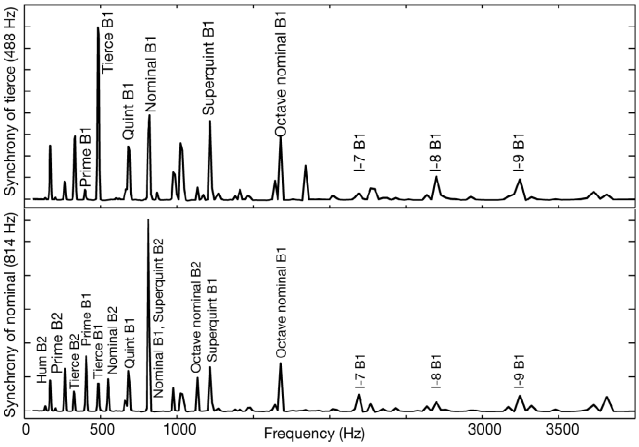

Fig. 4. Onset synchrony of two bell partials.

Matrix **O** contains many synchronicity patterns corresponding to the same bell, as each pattern represents onset synchrony of one of the bell's partials with all of the others. To discover all distinct groups of synchronicity patterns, we cluster the matrix with the intelligent k-means algorithm [18, pp. 93-96]. We use this algorithm, because it can automatically estimate the number of distinct clusters with the so-called "anomalous pattern approach". The rationale of anomalous patterns is that each cluster starts with the element furthest away from all of the others (anomalous pattern). Elements are then iteratively added to the cluster according to their distance to the cluster center. Authors suggest using Euclidean distances for finding anomalous patterns and forming clusters. Instead, we propose to use correlation, as it is better suited for comparing synchronicity patterns.

Algorithm 1: Finding clusters of synchronicity patterns

1. $t \leftarrow 0$

2. $\mathsf{S} \leftarrow$ { rows of $\sqrt{\mathbf{O}}$ (element-wise square root) }
3. $\mathsf{C}_t \leftarrow \varnothing$
   a. $\mathbf{c}_t \leftarrow$ element in S with the smallest average correlation to all other elements
   b. $\mathsf{C}_t \leftarrow \mathsf{C}_t \cup \{7\, \mathbf{s} \in \mathsf{S}$: correlation($\mathbf{s}$, $\mathbf{c}_t$) > T_c }
   c. $\mathbf{c}_t \leftarrow$ center of C_t
   d. if $\mathbf{c}_t$ changed in step *c*, go to step *b*
4. $\mathsf{S} \leftarrow \mathsf{S} \setminus \mathsf{C}_t$
5. $t \leftarrow t + 1$
6. if $|\mathsf{S}| > 1$, go to step *3*
7. remove all the found clusters C_t, where $|\mathsf{C}_t| = 1$
8. do regular k-means with $\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_K$ as initial centers

The proposed algorithm for estimating the number of clusters and their centers is as follows. S is first initialized to contain all synchronicity patterns, rescaled to correspond to partial magnitudes (2.). Each (anomalous) cluster starts with a synchronicity pattern furthest away from all other elements (3.a). Iteratively, similar-enough patterns (according to correlation with the cluster center) are added to the cluster (3.b), until no more patterns are found and the cluster is fully formed (3.d). The procedure is repeated until all patterns are clustered (6). Clusters with a single element are then removed (7) and regular k-means, initialized with the found centers, performed (8). The only parameter of Algorithm 1 is a threshold T_c, which controls the number of elements that will be iteratively added to the anomalous cluster in each iteration, and thus the size of such clusters. We evaluate the choice of T_c in section IV.

The obtained cluster centers represent groups of synchronicity patterns and are akin to spectra of bells in the analyzed recording. To assess their correspondence with the bell model given in eq. (2), we analyze each cluster center $\mathbf{c}_i$, and search for parameters of the bell model that best describe it, thus mapping each center to a particular bell sound. Specifically, for each center $\mathbf{c}_i$ we find bell model parameters *n* and *on* that maximize correlation between the bell model and $\mathbf{c}_i$:

$$\arg\max_{n,o} correlation(M_{n,o}, \mathbf{c}_i). \quad (5)$$

As the two bell model parameters *n* and *o* represent frequencies of the nominal and octave nominal partials, which are amongst the strongest bell partials (see TABLE I), we assume that they must be reflected in prominent peaks in the cluster center. Therefore, we constrain the search for *n* and *o* to all pairs of prominent cluster center peaks spaced in the range of 950 to 1500 cents, which corresponds to two standard deviations away from the mean (TABLE I). These constraints enable a fast exhaustive search for values of the two parameters that yield the highest correlation according to eq. (5). If two clusters are found to correspond to the same bell (their nominal frequencies are closer than 50 cents), they are merged.

The final outcome of the algorithm is a set of cluster centers; ideally, each center corresponds to one of the bells in a recording and indicates positions and magnitudes of the bell's partials (although the magnitudes are distorted due to flattening of the spectrogram and calculation of the delta spectrogram). We present an evaluation of the algorithm in section IV.

C. Non-negative Matrix Factorization

Decomposition of the time-frequency spectrogram into a set of basis vectors and their activations is performed by the multiplicative NMF algorithm based on the Kullback-Leibler I-Divergence. The algorithm incorporates sparsity and uncorrelatedness (orthogonality) constraints and has the following update rules for matrices **H** and **W** [19, p. 152]:

$$\mathbf{H} \leftarrow \mathbf{H} \cdot * \left(\mathbf{W}^T (\mathbf{V} \cdot / (\mathbf{WH})) - \alpha_{oH} (\mathbf{1}_{N \times N} - \mathbf{I}) \mathbf{H} \right)^{1+\alpha_{sH}}$$
$$\mathbf{W} \leftarrow \mathbf{W} \cdot * \left((\mathbf{V} \cdot / (\mathbf{WH})) \mathbf{H}^T - \alpha_{oW} \mathbf{W} (\mathbf{1}_{N \times N} - \mathbf{I}) \right)^{1+\alpha_{sW}}, \quad (6)$$

where $\cdot*$ and $\cdot/$ are the component-wise multiplication and division operators, α_{oH}, α_{oW} control the uncorrelatedness of **H** and **W** and may be used to reduce correlation between rows of **H** and columns of **W** (values should be >0), while α_{sH}, and α_{sW} control the sparsity of **H** and **W** (typical values are between 0.001 − 0.005 [19, p. 145]).

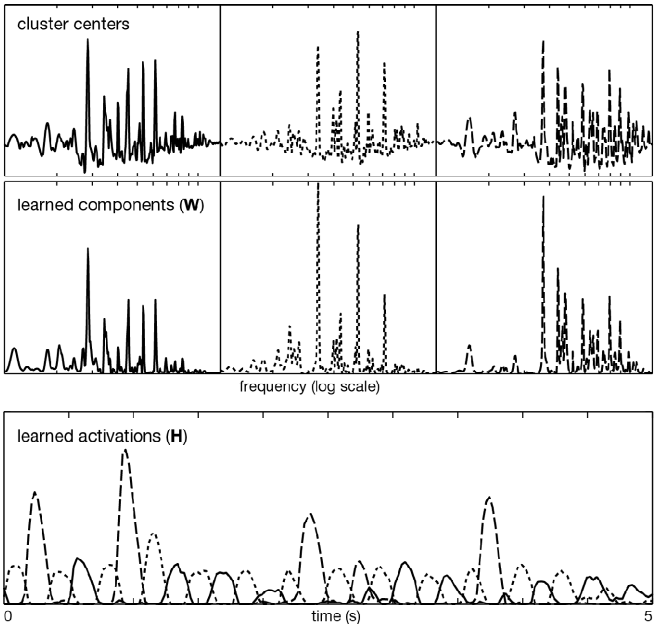

Fig. 5. Cluster centers and factorization of a bell chiming performance played on three church bells. Solid, dashed and dotted lines represent the three bells.

To obtain a meaningful decomposition, we initialize NMF with results of the algorithm outlined in section II.B. The number of found clusters determines the number of basis vectors and the matrix **W** is initialized so that its columns correspond to cluster centers scaled to the [0, 1] interval. The initial matrix **H** is obtained by non-negative least squares minimization [20] of the Frobenius norm ||**D**-**WH**||, where **D** represents the delta spectrogram. Factorization of the delta spectrogram **D** is then calculated by iterating the two steps of eq. (6). An example calculated on a bell chiming performance played on three church bells is shown in Fig. 5. Bell onsets are

clearly visible in the activation matrix **H** (bottom), which indicates that the decomposition is meaningful. The initial cluster centers (top) are similar to the obtained basis vectors (middle), however, the latter more faithfully represent the actual bell spectra, because cluster centers are derived from covariances of bell partials and are thus only approximate. Nominal frequencies of the three bells can be estimated from **W** by using eq. (5).

III. Transcription

To transcribe a bell chiming recording, we need to find which bells were played and when they were played - their onset times. Many current approaches to transcription, based on non-negative matrix factorization, use thresholding of the activation matrix **H** to find which notes were played and when [8-12]. For real-world signals, this is problematic, because decompositions are usually noisy. With church bell recordings, noise may be caused by a number of factors. Bell sounds decay for a long time and although bells are not usually played at the same time, the number of concurrently sounding bells (polyphony) is always high. Long decays accentuate interactions between bell partials: partials get amplified, cancelled or beat against each other, thus distorting values in the activation matrix. Partials also decay at different rates, so the spectrum of bells changes with time. This is a problem for factorization, where a constant spectrum is assumed for each bell; in our case, the problem is reduced by the use of the delta spectrogram. Bell chiming recordings are not synthetic; they are actual field recordings, not necessarily made by professionals and with professional equipment. Poor microphone placement or particular acoustics of church bell-towers may lead to dominance of certain bells or their partials; on the other hand, environmental noises, such as wind, or noises coming from the recording equipment, also introduce unnecessary artifacts in the activation matrix. Finally, recordings may contain fast passages that cause overlapping activations, leading to errors when thresholding is used for transcription; an example of two overlapping activations can be seen in the bottom part of near the end of the excerpt.

In this section, we propose a Markov chain model for transcription that integrates factorization and onset detection data with prior knowledge of bell chiming performance rules to find the most likely sequence of bells in a recording.

To build the model, bell onsets are first estimated with the complex domain onset detection function and peak picking algorithm [21]. The algorithm, which considers phase and amplitude changes to detect onsets, performs well with church bells, because of their sharp and percussive onsets.

Given N onset times and the fact that bells are seldom struck at the same time, transcription can be viewed as a problem of finding a sequence of states $x_1,x_2,x_3...x_N$ that best describes the analyzed signal; x_1 is placed at the first found onset, x_2 at the second and so on. x may represent a bell or a rest.

We formulate the problem as a random process - a first order time-inhomogeneous Markov chain, where the probability of the process occupying a certain state at a given time depends only on the preceding state:

$$P(X_{n+1}=x \mid X_1=x_1,\ldots,X_n=x_n)=P(X_{n+1}=x \mid X_n=x_n). \quad (7)$$

The state space S of the random variable X_n is defined as:

$$S=\{b_1,b_2,\ldots,b_M,r_1,r_2,\ldots,r_M\}, \quad (8)$$

where b_i denotes a bell (each bell is described by its corresponding spectral template in **W**), and r_i a rest.

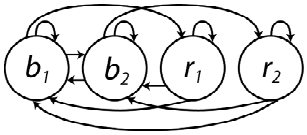

Fig. 6. The proposed Markov model for a recording with two bells.

In order to define the transition probability matrices $p_n(x|x_{n-1})=P(X_n=x \mid X_{n-1}=x_{n-1})$, we first introduce the scaled activity matrix **H'** as:

$$h'_{in}=\frac{h_{in}/a_n}{\max_n h_{in}/a_n}, \quad (9)$$

where a_n represents the low-pass filtered full-wave rectified signal waveform (at time n), which is used to reduce effects that changes in the overall signal level have on activations in **H**. The scaled values h'_{in} are taken to represent the probability of activation of bell i at time n.

The transition probability between two consecutive, but different bells is defined as:

$$p_n(b_i \mid b_j)=h'_{in}\, o_n\, t_n(b_i \mid b_j),\ i \neq j. \quad (10)$$

Next to h'_{in}, which accounts for the activation probability of bell i, o_n represents the onset probability, which is proportional to the value of the onset detection function at the onset n. $t_n(b_i|b_j)$ represents the probability of bell i following bell j in a bell chiming pattern played in the neighborhood of onset n; its calculation is explained further on in this section.

The probability of bell i being played twice in a row is calculated is a similar manner, whereby its previous activation h_{in-1} is subtracted from h_{in} as:

$$p_n(b_i \mid b_i)=(h_{in}-h_{in-1}\,\mathrm{G}(n,n-1,\sigma_s))o_n\, t_n(b_i \mid b_i). \quad (11)$$

G is the unnormalized Gaussian, centered at n-1, which models the time evolution of bell activations. As can be observed in Fig. 5, activations are roughly bell-shaped, and the Gaussian models their decay over time. Subtraction of bell activations improves transcription of fast passages, where activations may overlap and in cases when some bells are played considerably louder than others. Subtraction prevents assignment of consecutive onsets to the same bell and helps to reduce errors due to false positive onsets, which occur when two onsets are found for a bell instead of one.

Rests are used in our model to reduce effects of false positive onsets. Each bell b_i in the model has its corresponding rest state r_i (as shown in Fig. 6, transitions to other rest states are not allowed), to which it can transition to with probability:

$$p_n(r_i \mid b_i) = (1-o_n)\prod_{j=1}^{M}\left(1-h_{jn}\right). \quad (12)$$

A rest will thus be likely, if no bells are likely to be activated and the onset is also not likely. The above expression represents a conservative estimate of rest probability, as it requires that all of the bells, as well as the onset have low probabilities for a rest to be likely. However, as factorization of the delta spectrogram (upon which bell activation probabilities are based) corresponds closely to bell onsets, this is a valid choice, because high activations indicate onsets and not rests.

Transitions from rests to bells $p(b_i|r_j)$ are calculated according to eqs. (10) and (11), while the probability $p(r_i|r_i)$ of staying in the rest state is set to a constant value of 0.5.

We still need to define the calculation of $t_n(b_i|b_j)$, which denotes the probability of bell i following bell j in a bell chiming pattern played in the neighborhood of onset n. Here we consider the fact that bell chiming performances usually consist of gradually evolving rhythmic patterns. Therefore, each segment of a performance contains several repetitions of a single rhythmic pattern will small modifications. To calculate $t_n(b_i|b_j)$, the recording is first transcribed with $t_n(b_i|b_j)$ set to $1/M$. Bell transitions in the resulting transcription are counted and the frequencies of bell transitions are used to calculate the new $t_n(b_i|b_j)$. As rhythmic patterns evolve during a bell chiming performance, frequencies of transitions are counted within a window of size W_T around each onset n. The obtained values are noisy, because the transcription contains errors, but as each rhythmic pattern is repeated many times, errors are smoothed. The final transcription is calculated with the model that uses the calculated values for $t_n(b_i|b_j)$ in eqs. (10) and (11).

Given the described Markov model, the most likely sequence of bells and rests can be calculated efficiently with dynamic programming; the resulting set of onset times and bells represents the transcription of a recording.

IV. Evaluation and Discussion

We tested the algorithms on a set of field recordings from the digital archive of Slovenian folk music and dances EthnoMuse [2], which holds a large collection of bell chiming recordings, collected from the 1950s onwards. Recordings were made on various locations in Slovenia, with a variety of different equipment and recording setups, and may contain a considerable amount of noise, usually stemming from environmental factors, such as wind. Typically three to five bells were used in performances, some examples are given on http://lgm.fri.uni-lj.si/matic/TASLP2010.

A. Estimating the number of bells and their spectra

To test the algorithm described in section II.B, that estimates the number of bells and their approximate spectra in a recording, we collected a set of 30 bell chiming recordings performed on three to five church bells and manually labeled the nominal frequencies of bells used in performances (a total of 110 bells). Recordings were chosen randomly from field recordings containing bell chiming performances.

Basic parameters of the algorithm were not tuned specifically for the task and were set as follows. The magnitude spectrogram **F** was calculated with the constant-Q transform [22], using a maximum window size of 100ms, a step size of 25 ms and 60 frequency bins per octave, resulting in a good balance between time and frequency resolution. The spectrogram deltas **D** were calculated with a sliding window of N_d=4 frames (values between 2 and 6 all yield similar results) and covariance matrices on N_c=100 frames (2.5s) long segments. For the latter, windows from 2 to 8 seconds all produce similar results, while for windows shorter than 2 seconds the number of bell repetitions is too small to yield reliable covariance matrices.

For comparison, we calculated precision and recall of the correspondence of the found cluster centers to manual annotations for four variants of the algorithm with different time-frequency representations: deltas calculated on a flattened spectrogram (A_{fd}), deltas calculated on a normal spectrogram (A_{nd}), the flattened spectrogram (A_{fs}) and the normal spectrogram (A_{ns}). A cluster center was considered as correctly found, if its estimated nominal frequency was within 50 cents of the manual annotation. We used the Friedman test and Tukey's honestly significant difference test to estimate if differences between algorithm variants are significant with regard to the F_1 measure on individual recordings. Results at 95% confidence level are shown in Fig. 7. The y axis displays mean ranks of the algorithms (for each recording, algorithms are ranked according to their F_1 measure), so higher values mean better performance. The figure shows that the proposed approach with both spectral flattening and the delta spectrogram (A_{fd}) significantly outperforms other algorithm variants. Calculating the delta spectrogram improves results more than spectral flattening, although both are significant.

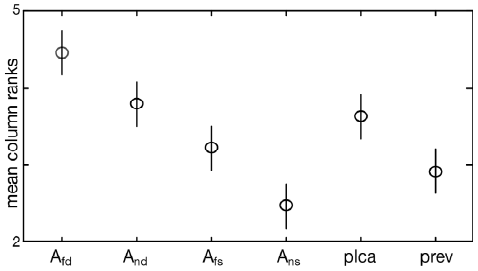

Fig. 7. Statistical evaluation of four variants of the proposed algorithm is shown: with and without spectral flattening, as well as with and without the delta spectrogram. Results of the PLCA algorithm and our previous approach are also shown.

We also assessed the effect of the threshold T_c on clustering of synchronicity patterns by evaluating the four algorithm variants with a range of T_c values (from 0.5 to 0.9). The average F_1 measures are shown in Fig. 8. As can be observed, spectral flattening provides a consistent improvement over the regular spectrogram. As weaker partials get accentuated, they also become more visible in synchronicity patterns, and the

fact that more bell partials are discovered makes the clustering process and bell model fitting more accurate.

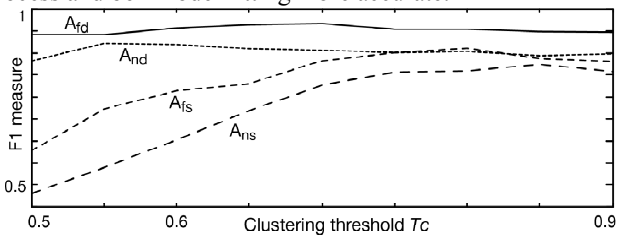

Fig. 8. F_1 measures for different values of the clustering threshold T_c, using the four algorithm variants.

Both variants that use the delta spectrogram are much less sensitive to the choice of the clustering threshold T_c, which indicates that correlations between synchronicity patterns obtained from the delta spectrogram are lower and thus the patterns are more distinct for different bells than patterns derived from the spectrogram. We attribute the difference mainly to the fact that bell sounds have very long decay times, so in a given recording, most of the bells are sounding most of the time. When we consider that bells have very rich inharmonic spectra and frequently exhibit beating due to their imperfect shape, the spectrogram of a mixture of bells is difficult to analyze directly. Synchronicity patterns, as well as clusters, derived from the spectrogram, are not well separated; a single pattern contains partials of several bells and a single cluster contains patterns of several bells. Thus, correlations between cluster centers are higher and the algorithm is more sensitive to the choice of the clustering threshold T_c. The delta spectrogram, on the other hand, reduces the amount of information in the time-frequency representation, consequently decorrelates synchronicity patterns and improves the formation of clusters. The obtained results also indicate that factorizations derived from the delta spectrogram will be more meaningful (accurate), leading to more accurate transcriptions. Since most of the authors that use NMF for transcription decompose time-frequency spectrograms directly, it would be interesting to observe whether factorizing the delta spectrogram would also benefit their approaches; we speculate that it could improve the accuracy of systems that transcribe music containing instruments with pronounced onsets and longer decays, such as the piano.

We also compared the proposed algorithm to two other approaches; results are shown in Fig. 7. First, we compared it to our previous bell chiming transcription algorithm [23]. The algorithm uses NMF with selective sparsity constraints and adaptation of the number of basis vectors to factorize the time-frequency spectrogram. Its performance suffers mostly due to problems with factorization of the time-frequency representation, as it is not using the delta spectrogram representation.

We also tested how probabilistic latent component analysis (PLCA) [24] can be used for estimation of spectral templates. PLCA is related to NMF; it recasts factorization in a probabilistic framework as:

$$\mathbf{V} \cong \mathbf{WZH}, \tag{13}$$

where columns of **W** and rows of **H** are multinomial probability distributions and **Z** a diagonal matrix of mixing weights. For our experiments, we used the PLCA variant introduced by Weiss et al. [25] for identification of repeated patterns in music. Their algorithm allows for sparsity constraints over **W**, **H** and **Z**, where constraints on the latter can be used to iteratively learn the number of components (spectral templates) in the mixture. Therefore, learning can be started with many components and the sparsity constraints on **Z** prune out components that do not contribute significantly to the reconstruction of **V**. We used the PLCA algorithm to decompose the delta spectrogram **D**.

Although all compared algorithms perform reasonably well, the proposed approach is the most accurate, and the difference is statistically significant. Better performance is important, as errors propagate and cause false negatives (missed bell spectral templates) and false positives (extraneous templates) in transcription. Advantages in comparison to PLCA are twofold. First, the proposed algorithm uses a local approach to find synchronicity patterns; covariance matrices are calculated on short segments of the entire recording and then combined based on magnitudes of partials in these segments. On the other hand, PLCA works globally by iteratively minimizing the factorization error. The difference is important when searching for bells that are played quietly or are not frequently played. PLCA tends to ignore such bells and rather focuses on minimizing the overall error which may lie in varying amplitude envelopes of bell partials or noise. The local nature of our approach does not fail for such cases, as the bells stand out in individual local segments and consequently also show in the global onset synchrony matrix **O**. Another advantage of the proposed approach is its use of prior knowledge encoded in the model of bell partial positions, presented in section II.A. Namely, false positives are reduced by grouping clusters of synchronicity patterns representing the same bell.

Analysis of errors made by the proposed algorithm showed that false negatives (missed bells) mostly occur with bells that are played very quietly in the background, so their onsets, which should be apparent in the delta spectrogram, are lost between other louder bells. Another source of false negatives and false positives, are incorrectly estimated nominal frequencies of the found bell templates, leading either to incorrect merging of the found clusters or to incorrect labeling of bells.

B. Transcription

To test the transcription algorithm, we manually transcribed beginnings of 15 bell chiming performances containing a total of 1123 bell onsets. Recordings were randomly chosen from the test set of the bell finding algorithm. Six excerpts and their transcriptions can be found at: http://lgm.fri.uni-lj.si/matic/TASLP2010.

Parameters of the algorithm were set as follows. NMF was calculated without sparsity constraints, but with uncorrelatedness constraints on the activity matrix **H** ($\alpha_{oH} = 1$, $\alpha_{oW} = \alpha_{sH,} = \alpha_{sW} = 0$). These values were determined experimentally, however they do not have a major effect on

transcription accuracy. Intuitively, activations of different bells in **H** should not be highly correlated, as bells are usually not struck at the same time; adding the uncorrelatedness constraint on **H** encodes this knowledge into factorization. The width of bell activations σ_s was set to 150 ms, corresponding to the average width of activations of several analyzed bells, and the width of the window for calculating transition probabilities W_T to 30 seconds, corresponding to the average length of bell chiming patterns.

To evaluate how various choices made when designing our transcription algorithm influence its performance, we tested several variants of the algorithm: *A* – the described algorithm, *B* – excluding bell transition probabilities $t_n(b_i|b_j)$, thus making all bell transitions equally probable, *C* – excluding bell transition probabilities and subtraction of bell activations and *D* – using the annotated onsets instead of the calculated ones. Since we are not aware of other algorithms for transcription of recordings containing church bells, we compared the proposed approach to an approach that uses simple thresholding of the activity matrix **H** for transcription (*E*) and to our previously published algorithm [23] (*F*). For all variants, we calculated precision and recall of transcriptions, where bells were considered as correctly transcribed, if their onsets were within 50 ms of the manually annotated onset. To compare the algorithms, we used the Friedman test combined with Tukey's honestly significant difference test to estimate whether differences are significant with regard to the F_1 measure on individual examples. Results at 95% confidence level are shown in Fig. 9, and summarized in TABLE II. The second column of the table contains average precision and recall of transcriptions and the third column average precision and recall of onsets of the found notes (not considering bells assigned to the onsets).

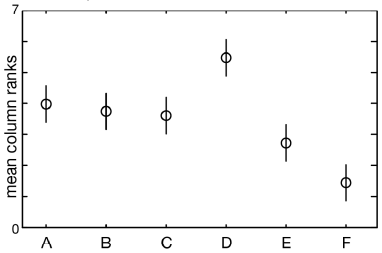

Fig. 9. Comparison of variants of the proposed transcription algorithm (*A*-*D*), thresholding of **H** (*E*) and our previous approach (*F*).

TABLE II
EVALUATION OF THE TRANSCRIPTION ALGORITHM.

algorithm	transcription precision/recall	onset precision/recall
A: proposed algorithm	0.90 / 0.91	0.93 / 0.94
B: no transition prob.	0.88 / 0.9	0.93 / 0.95
C: no act. subtraction	0.87 / 0.91	0.92 / 0.96
D: perfect onsets	0.95 / 0.95	1 / 1
E: thresholding of H	0.85 / 0.92	0.89 / 0.97
F: our previous algorithm	0.74 / 0.83	0.87 / 0.97

Results show that differences between variants of our algorithm (*A*-*C*) are not very big, and although *A* is the most accurate, differences to *B* and *C* are not statistically significant. However, improvements introduced by the subtraction of bell activations and bell transition probabilities make *A* significantly more accurate than the simple thresholding method (*E*). Transition probabilities are useful for identification of bells played in fast passages or played quietly, where they help to resolve ambiguous overlapping activations. An example is shown in Fig. 10. In the section pointed to by "potential error 2", two bells are played in close proximity. Their order is not discernable from activations in **H**, as they overlap almost perfectly, but if transition probabilities are used, the correct order (the one that is also dominant in the current pattern) will be transcribed. "Potential error 1" in Fig. 10 shows an onset detection error (false positive), which is resolved if subtraction of bell activations is used. The false positive onset is transcribed as a rest, because the bell's activation is subtracted at the second (false positive) onset.

As (*D*) shows, approx. half of the errors are due to incorrect onset detection (error 3 in Fig. 10 shows a missed onset), and precision, as well as recall, are raised by approx. 0.05 if the annotated onsets are used (the difference is statistically significant).

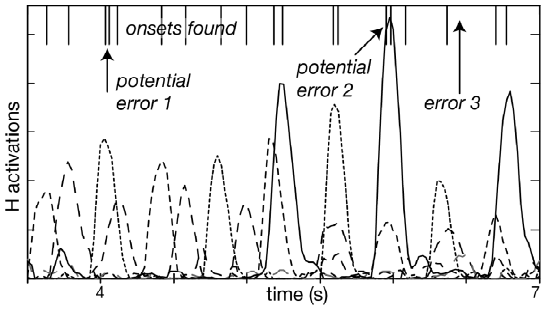

Fig. 10. Two types of errors, which can be resolved by using the bell transition probabilities and subtraction of bell activations, are shown. An onset-related error is also shown. Activations in the matrix **H** are plotted with different dashed lines for different basis vectors (bells), while onsets found by the onset detection algorithm are shown as vertical lines in the top part of the figure.

The proposed algorithm (*A*) significantly outperforms the method that uses simple thresholding of **H** (*E*), but it mostly improves precision and not recall. Thus, it mostly seems to be filtering out false positive onsets, but it does not improve the labeling of correctly found onsets. This indicates that the remaining 5% of errors made by the algorithm (if we ignore onset-related errors) occur with bells that are very difficult to discern from the time-frequency representation, either because they are very quietly played or because of strong interactions between their partials. The latter may be the result of beating or other factors such as poor microphone placement, weather conditions and bell tower acoustics. Finally, the accuracy of our previous algorithm (*F*) mostly suffers because of incorrectly identified bells.

V. CONCLUSION

In the paper, we describe an algorithm for transcription of recordings of bell chiming performances and other similar

music played on church bells (i.e. change ringing). The strengths of the proposed approach stem from its use of the delta spectrogram for time-frequency representation, clustering and previous knowledge of church bell acoustics for determining the number of bells and their approximate spectra, and a probabilistic framework for transcription. Its accuracy on a real-world collection of Slovenian bell chiming field recordings is at a good 97% precision and recall on the bell finding task and at 90% precision and recall on transcription. Such performance is already sufficient for practical applications and the algorithm is currently being used by researchers at the Institute of Ethnomusicology for analysis of recordings in their EthnoMuse archive. The validity of the proposed approach is further supported by its successful application to transcription of melodies played by bell-playing clocks [26]. With some modifications of the acoustical model and the transcription framework, we obtained accurate transcriptions (over 90% precision and recall) of recordings from the Nederlandse Liederenbank, where we are currently testing the algorithm within a melody-based retrieval framework.

There is room for improvement, onset detection and better handling of beating being two directions, but our first future goal is automatic extraction of characteristic bell chiming patterns, which is of interest to researchers and performers. We also plan to integrate the algorithm into the EthnoMuse platform, where it will become part of a retrieval system that will support queries based on bell chiming patterns and recording excerpts.

REFERENCES

[1] M. Kovačič "New Contexts, Esthetics, and Transfer in Bell-chiming Tradition," *Slovene studies,* vol. 29, pp. 19-34, 2007.

[2] M. Marolt, *et al.*, "Ethnomuse: Archiving Folk Music and Dance Culture," in *Eurocon 2009*, St. Petersburg, Russia, 2009.

[3] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.

[4] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 1066-1074, 2007.

[5] A. Cont, "Realtime Audio to Score Alignment for Polyphonic Music Instruments, using Sparse Non-Negative Constraints and Hierarchical HMMS," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006.*, Toulouse, France., 2006.

[6] S. A. Abdallah and M. D. Plumbley, "Unsupervised analysis of polyphonic music by sparse coding," *Neural Networks, IEEE Transactions on,* vol. 17, pp. 179-196, 2006.

[7] 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.*, 2003, pp. 177-180.

[8] N. Bertin, *et al.*, "Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Polyphonic Music Transcription," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, pp. 538-549, 2010.

[9] E. Vincent, *et al.*, "Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 109-112.

[10] S. A. Raczynski, *et al.*, "Multipitch Analysis with Harmonic Nonnegative Matrix Approximation," in *ISMIR 2007, 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007, pp. 381-386.

[11] B. Niedermayer, "Non-negative Matrix Division for the Automatic Transcription of Polyphonic Music," in *ISMIR 2008, 9th International Conference on Music Information Retrieval*, Philadelphia, USA, 2008, pp. 544-545.

[12] G. Grindlay and D. P. W. Ellis, "A Probabilistic Subspace Model for Multi-Instrument Polyphonic Transcription," in *11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, 2010, pp. 21-26.

[13] E. Vincent, *et al.*, "Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008*, Las Vegas, USA, 2008, pp. 109-112.

[14] W. A. Hibbert, "The Quantification of Strike Pitch and Pitch Shifts in Church Bells," Ph.D., Faculty of Mathematics, Computing and Technology, The Open University, Milton Keynes, UK, 2008.

[15] E. Vincent and M. D. Plumbley, "Low Bit-Rate Object Coding of Musical Audio Using Bayesian Harmonic Models," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 1273-1282, 2007.

[16] A. Klapuri, "A perceptually motivated multiple-FO estimation method for polyphonic music signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2005.

[17] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, 2004.

[18] B. Mirkin, *Clustering For Data Mining: A Data Recovery Approach*: Chapman & Hall/CRC, 2005.

[19] A. Cichocki, *et al.*, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Chichester, UK: John Wiley & Sons Ltd, 2009.

[20] M. H. Van Benthem and M. R. Keenan, "Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems," *Journal of Chemometrics,* vol. 18, pp. 441-450, 2004.

[21] J. P. Bello, *et al.*, "On the use of phase and energy for musical onset detection in the complex domain," *Signal Processing Letters, IEEE,* vol. 11, pp. 553-556, 2004.

[22] J. C. Brown, "Calculation of a Constant-Q Spectral Transform," *Journal of the Acoustical Society of America,* vol. 89, pp. 425-434, Jan 1991.

[23] M. Marolt, "Non-negative matrix factorization with selective sparsity constraints for transcription of bell chiming recordings," in *Sound and Music Computing Conference*, Porto, Portugal, 2009, pp. 137-142.

[24] P. Smaragdis, *et al.*, "Sparse and shift-invariant feature extraction from non-negative data," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 2069-2072.

[25] R. J. Weiss and J. P. Bello, "Identifying Repeated Patterns in Music Using Sparse Convolutive Non-negative Matrix Factorization," in *11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, 2010.

[26] M. Marolt and M. Lefeber, "It's Time for a Song – Transcribing Recordings of Bell-playing Clocks," in *ISMIR, 11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, 2010, pp. 333-338.

Matija Marolt (M'96) received the B.S. and Ph.D. degrees, both in computer science, from University of Ljubljana, Faculty of Computer and Information Science, Slovenia in 1995 and 2002 respectively.

He is assistant professor with Faculty of Computer and Information Science, University of Ljubljana, where he has been working since 1995. He is a member of Laboratory of Computer Graphics and Multimedia, where his research interests include music information retrieval, specifically semantic description of audio, retrieval and organization of ethnomusicological archives and audio-visual interaction.