AUTOMATIC TRANSCRIPTION OF PITCHED AND UNPITCHED SOUNDS FROM POLYPHONIC MUSIC

Emmanouil Benetos[⋆] *Sebastian Ewert*[†] *Tillman Weyde*[⋆]

[⋆] Department of Computer Science, City University London, London, UK
[†] Centre for Digital Music, Queen Mary University of London, London, UK

ABSTRACT

Automatic transcription of polyphonic music has been an active research field for several years and is considered by many to be a key enabling technology in music signal processing. However, current transcription approaches either focus on detecting pitched sounds (from pitched musical instruments) or on detecting unpitched sounds (from drum kits). In this paper, we propose a method that jointly transcribes pitched and unpitched sounds from polyphonic music recordings. The proposed model extends the probabilistic latent component analysis algorithm and supports the detection of pitched sounds from multiple instruments as well as the detection of unpitched sounds from drum kit components, including bass drums, snare drums, cymbals, hi-hats, and toms. Our experiments based on polyphonic Western music containing both pitched and unpitched instruments led to very encouraging results in multi-pitch detection and drum transcription tasks.

Index Terms— Music signal analysis, automatic music transcription, multi-pitch detection, drum transcription

1. INTRODUCTION

Automatic music transcription refers to the process of converting an acoustic musical signal into some form of music notation, and is considered to be a key problem in the field of music signal processing, having several applications in music information retrieval, interactive music systems, and computational musicology [1]. However, the area of automatic transcription is split into two strands, with one focusing on transcription of pitched sounds (i.e. multi-pitch detection) and the other on transcribing unpitched sounds (typically drum sounds). Even though research is active in both topics, currently no attempt has been made to jointly transcribe pitched and unpitched musical instruments, even though a large subset of recorded music contains instances of both (e.g. pop, rock, jazz).

Regarding automatic transcription of harmonic sounds, a large subset of current approaches employs spectrogram factorization techniques [2], such as non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA). Related work includes the PLCA-based system of Grindlay and Ellis [3], which supports multiple spectral templates for each pitch and instrument source and models fixed spectral templates as a linear combination of basic instrument models. Also, Fuentes et al. [4] proposed a pitched transcription system based on PLCA, which decomposes an input music signal into a harmonic component and a noise component. The harmonic signal represents each note as a weighted sum of narrowband log-spectra which are also shifted across log-frequency. Finally, the authors in [5] proposed a PLCA-based pitched sound transcription model which supports multiple templates per pitch and instrument, and also uses pre-shifted and pre-extracted templates across log frequency for supporting tuning changes and frequency modulations (the model of [5] is used as the pitched component for the proposed model).

Related work on drum transcription includes the system of Lindsay-Smith et al. [6], who employ convolutive NMF for transcribing solo drum loops and represent each drum template as a time-frequency patch. Gillet and Richard [7] proposed a system for transcribing and separating drums from polyphonic music signals, using a harmonic/noise decomposition and Wiener filtering-based separation. Finally, Paulus and Klapuri [8] transcribed drum sounds using a network of connected hidden Markov models, using the ENST drums dataset for evaluation.

In this work, we propose a novel system for joint transcription of pitched and unpitched sounds from polyphonic music. To the authors' knowledge, this is the first transcription system to perform joint transcription of multiple pitches and drum sounds. The model extends the PLCA algorithm [2] and decomposes an input log-frequency spectrogram into a pitched and an unpitched component. The pitched component supports multiple-instrument polyphonic music, as well as tuning changes and frequency modulations. The unpitched component supports the detection of overlapping sounds from drum kit instruments (bass drum, snare drum, hi-hats, cymbals, toms). For evaluation, we use a recordings from the TRIOS [9] and RWC [10] databases which contain pitched and unpitched sounds. For the RWC data, we also create temporally aligned ground truth by applying the music synchronization algorithm of [11] to the non-aligned annotations found in the database. A good level of accuracy on multi-pitch detection and drum transcription is reported on complex polyphonic recordings containing pitched and unpitched sounds.

The outline of this paper is as follows. In Section 2, the proposed transcription model is presented, along with the algorithm for parameter estimation and the postprocessing procedure. Section 3 describes the datasets used for training and testing, the evaluation metrics, and the experimental results. Finally, conclusions are drawn and future directions are indicated in Section 4.

2. PROPOSED METHOD

In the following, we describe a method for the automatic transcription of both pitched and unpitched sounds from polyphonic Western music. The proposed model will be able to detect multiple pitches produced by multiple instruments, using several spectral templates per pitch and instrument source. Tuning changes and frequency modulations will be supported by incorporating shift-invariance

Emmanouil Benetos is supported by a City University London Research Fellowship.

across log-frequency. In addition, the model will be able to detect and classify unpitched sounds produced by drum kit components, including bass drums, snare drums, hi-hats, cymbals, and toms.

2.1. Model

The proposed model extends the transcription model for detecting pitched sounds introduced in [5], which was based on probabilistic latent component analysis (PLCA) [2] and used pre-extracted note templates from multiple harmonic instruments. In the following, we extend the approach of [5] by incorporating an additional unpitched component that adds the ability to model the various instruments in a drum kit. The proposed model uses as input a normalised log-frequency spectrogram and decomposes it as a pitched component (which is modelled according to [5]) and an unpitched component, supporting several drum kit instruments.

The model approximates the input log-spectrogram $V_{\omega,t}$ (where ω stands for log-frequency and t stands for time) as a bivariate probability distribution $P(\omega, t)$, which is factored as:

$$P(\omega, t) = P(t)P(\omega|t) \tag{1}$$

where $P(t)$ is the frame probability (known quantity) and $P(\omega|t)$ is the conditional distribution over log-frequency bins. $P(\omega|t)$ is further decomposed as a pitched and unpitched component:

$$P(\omega|t) = P(r = h|t)P_h(\omega|t) + P(r = u|t)P_u(\omega|t) \tag{2}$$

where $P_h(\omega|t)$ is the spectrogram approximation for the pitched component of the signal and $P_u(\omega|t)$ is the approximation for the unpitched component. The probability $P(r|t)$ ($r \in \{h, u\}$) corresponds to the weights of the pitched and unpitched components over time.

The pitched component is decomposed as:

$$P_h(\omega|t) = \sum_{p,f,s} P_h(\omega|s, p, f)P_h(f|p, t)P_h(s|p, t)P_h(p|t) \tag{3}$$

where $p \in \{21, \ldots, 108\}$ denotes pitch in MIDI scale, s denotes the pitched instrument index, and f is the shifting parameter across log-frequency, denoting small pitch changes. $P_h(\omega|s, p, f)$ are the log-spectral templates per pitch p and instrument s, which are also shifted across log-frequency according to parameter f. Our time-frequency representation has a spectral resolution of 5 bins per semitone and, by constraining parameter f to $f \in \{1, \ldots, 5\}$, the spectral templates can be shifted by ± 0.5 semitones (thus, $f = 3$ denotes the ideal tuning position). $P_h(f|p, t)$ is the time-varying log-frequency shifting per pitch, $P_h(s|p, t)$ denotes the instrument contribution per pitch over time (useful for instrument assignment evaluation), and finally $P_h(p|t)$ is the pitch activation, which is used to evaluate the model for multi-pitch detection.

The unpitched component is decomposed as:

$$P_u(\omega|t) = \sum_{d,z} P_u(\omega|d, z)P_u(d|t)P_u(z|d, t) \tag{4}$$

where d denotes the drum kit component (in this paper, it can be bass drum, snare drum, hi-hat, cymbals, or toms) and z is the index for the 'exemplars' that are used for each component. Thus, $P_u(\omega|d, z)$ denotes the z-th log-spectral template for drum component d, $P_u(d|t)$ is the drum component activation (used for drum transcription evaluation), and finally $P_u(z|d, t)$ is the exemplar contribution per drum component over time.

2.2. Parameter Estimation

The unknown parameters in the model are $P(r|t)$, $P_h(f|p, t)$, $P_h(s|p, t)$, $P_h(p|t)$, $P_u(d|t)$, and $P_u(z|d, t)$. The pitched and unpitched templates ($P_h(\omega|s, p, f)$ and $P_u(\omega|d, z)$, respectively) are pre-extracted and thus remain fixed.

In order to estimate unknown model parameters, we use the expectation-maximization (EM) algorithm [12]. Given the input log-frequency spectrogram $V_{\omega,t}$, the model log-likelihood is given by:

$$\mathcal{L} = \sum_{\omega,t} V_{\omega,t} \log\big(P(\omega, t)\big). \tag{5}$$

In the *Expectation* step, the posterior distribution over the hidden variables (p, s, f, d, z) is calculated using Bayes' theorem:

$$P(s, p, f, r = h|\omega, t) = \frac{P(r = h|t)P_h(\omega|s, p, f)P_h(f|p, t)P_h(s|p, t)P_h(p|t)}{P(\omega|t)} \tag{6}$$

$$P(d, z, r = u|\omega, t) = \frac{P(r = u|t)P_u(d|t)P_u(z|d, t)}{P(\omega|t)} \tag{7}$$

For the *Maximization* step, we utilise the posteriors of (6-7) for maximizing the log-likelihood of (5), resulting in the following update equations for the pitched components:

$$P(r = h|t) \propto \sum_{s,p,f,\omega} V_{\omega,t}P(s, p, f, r = h|\omega, t) \tag{8}$$

$$P_h(f|p, t) = \frac{\sum_{\omega,s} V_{\omega,t}P(s, p, f, r = h|\omega, t)}{\sum_{\omega,s,f} V_{\omega,t}P(s, p, f, r = h|\omega, t)} \tag{9}$$

$$P_h(s|p, t) = \frac{\sum_{f,\omega} V_{\omega,t}P(s, p, f, r = h|\omega, t)}{\sum_{f,\omega,s} V_{\omega,t}P(s, p, f, r = h|\omega, t)} \tag{10}$$

$$P_h(p|t) = \frac{\sum_{s,f,\omega} V_{\omega,t}P(s, p, f, r = h|\omega, t)}{\sum_{s,f,\omega,p} V_{\omega,t}P(s, p, f, r = h|\omega, t)} \tag{11}$$

The update equations for the unpitched components of the model are as follows:

$$P(r = u|t) \propto \sum_{d,z,\omega} V_{\omega,t}P(d, z, r = u|\omega, t) \tag{12}$$

$$P_u(d|t) = \frac{\sum_{z,\omega} V_{\omega,t}P(d, z, r = u|\omega, t)}{\sum_{z,\omega,d} V_{\omega,t}P(d, z, r = u|\omega, t)} \tag{13}$$

$$P_u(z|d, t) = \frac{\sum_{\omega} V_{\omega,t}P(d, z, r = u|\omega, t)}{\sum_{\omega,z} V_{\omega,t}P(d, z, r = u|\omega, t)} \tag{14}$$

Eqs. (8) and (12) are normalised by the sum of their respective numerators, i.e. $\sum_{s,p,f,\omega} V_{\omega,t}P(s, p, f, r = h|\omega, t) + \sum_{d,z,\omega} V_{\omega,t}P(d, z, r = u|\omega, t)$. For estimating the unknown parameters, eqs. (8)-(14) are iterated until convergence. By keeping the spectral templates $P_h(\omega|s, p, f)$ and $P_u(\omega|d, z)$ fixed, the model required about 20-30 iterations for convergence.

Since typically in polyphonic music only few notes are active at a given time frame and that few instruments are responsible for producing a specific note at a time frame, we also impose sparsity constraints on model parameters. In specific, we impose sparsity constraints on the pitched component through $P_h(p|t)$ and $P_h(s|p, t)$, as well as on the unpitched component through $P_u(z|d, t)$. The motivation for imposing sparsity on $P_u(z|d, t)$ is for not allowing combinations of many drum exemplars to approximate an input unpitched

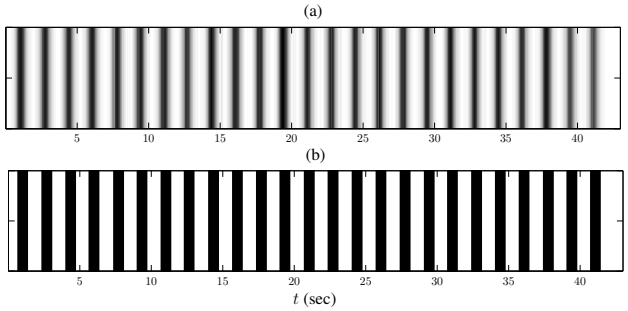

Fig. 1. Bass drum transcription for the 'Take Five' recording, using templates from the same source. (a) The transcription probability $P_u(d = bd|t)$, where bd denotes the bass drum. (b) The respective ground truth.

sound, as the model itself is very rich. The aforementioned constraints are incorporated similarly to the method described in [13], by modifying update equations (10), (11), and (14), by setting the numerators and denominators to a power greater than 1, thus sharpening the probability distributions. In this work, the sparsity parameter for the aforementioned distributions is set to 1.1.

2.3. Postprocessing

The resulting MIDI-scale transcription for the pitched component is given by:

$$P_h(p,t) = P(t)P(r = h|t)P_h(p|t) \tag{15}$$

The pitched component of the model can also output a high-resolution time-pitch representation by exploiting information from the pitch shifting parameter $P_h(f|p,t)$:

$$P_h(f,p,t) = P(t)P(r = h|t)P(f|p,t)P(p|t) \tag{16}$$

By stacking together slices of $P_h(f,p,t)$ for all p, a time-pitch representation $P_h(f',t)$ with a spectral resolution equivalent to the resolution of the input time-frequency representation can be created, which is useful for visualising pitch content for extracting tuning information.

In addition, the resulting drum transcription (using the unpitched model components) is given by:

$$P_u(d,t) = P(t)P(r = u|t)P(d|t) \tag{17}$$

In order to derive a binary piano-roll and "drum-roll" representation, a post-processing procedure is used to analyse the pitched and unpitched transcriptions in (15) and (17). As in the vast majority of automatic transcription systems using spectrogram factorization techniques (e.g. [14, 13]), we perform thresholding on the transcriptions. For the pitched transcription, we additionally remove detected events with a duration less than 80ms; short events are not removed from the unpitched transcription due to the percussive nature of drum sounds.

For example, in Fig. 1, the bass drum transcription for the 'Take Five' recording used in the evaluations can be seen, along with the ground truth. In Fig. 2, the pitched transcription of the same recording can be seen; even though the note durations are not well estimated, there are very few spurious notes and note onsets are for the most part at their correct positions.

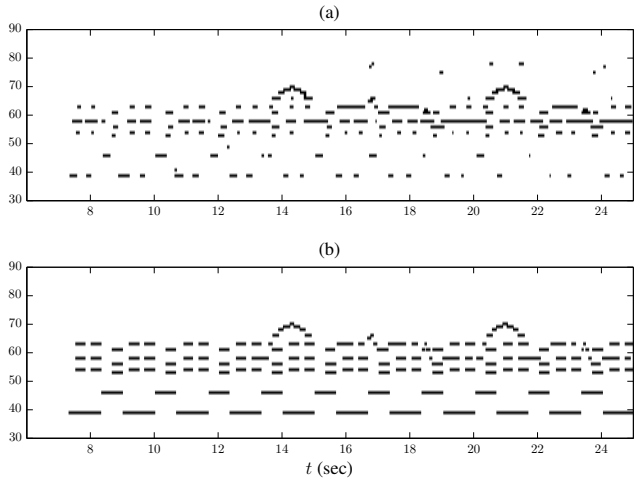

Fig. 2. Pitched transcription for the 'Take Five' recording. (a) The piano-roll transcription. (b) The respective ground-truth.

3. EVALUATION

3.1. Training Data

For training the system with pre-extracted templates of pitched sounds, we used isolated note samples of piano, saxophone, and double bass, taken from the RWC and MAPS databases [15, 10]. The complete note range of the instruments is used, given the available training data. For the pre-extracted drum templates, we used isolated drum sounds from bass drums (7 recordings), snare drums (43 recordings), hi-hats (25 recordings), cymbals (28 recordings), and toms (33 recordings), taken from the RWC database [10]. In addition, we extracted templates for bass drum, snare drum, and cymbals using individual tracks of the multi-track 'Take Five' recording from the TRIOS dataset [9], which is used in some of our experiments. As a time-frequency representation, we use the constant-Q transform with spectral resolution of 60 bins/octave [16]. Pitched templates are computed using the PLCA algorithm with one component [2]. Due to the transient nature of drum sounds, we simply added exemplars directly to the dictionary $P_u(\omega|d,z)$, by sampling at the CQT spectrograms with a 40ms step.

3.2. Test Data

For testing, we used the complete mix of the 'Take Five' recording from the TRIOS dataset [9], which additionally contains manually-aligned MIDI ground truth for the piano, saxophone, and drum tracks (it is the only recording in the dataset that also contains drums). For comparative purposes, we evaluated the recording using drum templates from the RWC database and also from drum templates directly extracted from the drum tracks of the recording (as to show the potential upper limit in drum transcription using the proposed method).

We conducted additional experiments using five pieces taken from the RWC Jazz database [10], which comprise both harmonic and percussive instruments (pieces RWC-MDB-J 16-20). While MIDI files provided by the RWC database encode the notes and instruments played in each piece, they are not temporally aligned with the audio and thus do not show when the notes are played. To generate a ground truth transcription from these MIDI files, we employ a high-resolution music synchronization approach described

	F_{mp}	F_{bd}	F_{sd}	F_{cym}
RWC drums	77.47%	29.51%	48.18%	49.37%
RWC + TRIOS drums	77.18%	92.00%	57.52%	60.76%

Table 1. Transcription results for the 'Take Five' recording from the TRIOS dataset, using drum templates from the RWC database only and the RWC + TRIOS databases.

	F_{mp}	F_{bd}	F_{sd}	F_{hh}	F_{cym}	F_{dr}
50ms	37.42%	20.06%	27.27%	60.81%	35.36%	35.88%
100ms	47.98%	26.26%	40.09%	68.15%	52.56%	46.77%

Table 2. Average transcription results for the 5 RWC recordings, using 50ms and 100ms onset tolerance for evaluation.

in [11]. The procedure is based on Dynamic Time Warping (DTW) and chroma features but extends previous synchronization methods by introducing onset-based features to yield a higher alignment accuracy. Using the resulting alignment we determine for each note event the corresponding position in the audio and update its onset position and duration in the MIDI file accordingly. Due to certain mis-alignments in sections where only drums are present, we evaluated the first two minutes of all five recordings.

3.3. Metrics

We evaluate the performance of the proposed system for multi-pitch detection and drum transcription, using onset-based metrics which are used in the MIREX note tracking evaluations [17]. For multi-pitch detection, a detected note is considered correct if its pitch matches a ground truth pitch and its onset is within a 50ms tolerance of a ground-truth onset. For drum transcription, a drum event is considered correct if it belongs to the correct drum kit component and its onset is within a 50ms tolerance of a ground-truth onset. Duplicates found within the same tolerance interval are considered false alarms. Since the ground truth generated using the automatic alignment procedure is not as precise as manually generated annotations, we additionally conducted comparative experiments on the RWC test data using a slightly increased tolerance of 100ms.

As evaluation metrics, we use the onset-based precision, recall, and F-measure (defined e.g. in [18]). In the following, the F-measure for multi-pitch detection is denoted as F_{mp}, whereas the F-measure for the bass drum, snare drum, hi-hat, and cymbals is denoted by F_{bd}, F_{sd}, F_{hh}, and F_{cym}, respectively. Also, an average drum transcription metric is used, namely F_{dr}, averaging all metrics for the drum components. It should be noted that the test recordings do not contain sounds from toms, although the proposed system does contain and support tom templates.

3.4. Results

Transcription results using the 'Take Five' recording from the TRIOS dataset are shown in Table 1. It can be seen that the drum transcription performance using templates from the same recording can radically increase performance, as is especially evident for the bass drum. By examining the spectral shape of the templates, it can be seen that the RWC templates for the bass drum are tuned much higher. Irrespective of the drum dictionary, the pitched transcription performance remains the same. For multi-pitch detection the precision (81.88%) is much higher than the recall (72.98%), indicating that the proposed method has more issues with missed detections than with false alarms.

Results using the 5 piano, bass, and drum recordings from the RWC database are shown in Table 2. The performance in both multi-pitch detection and drum transcription is much lower compared to the recording taken from the TRIOS dataset, which can be partly attributed to inaccuracies in the ground truth resulting from the use of an automatic alignment procedure. Other reasons include the much more complex nature of the pieces, with rapid piano playing and multiple overlapping drum sounds. However, it is worth noting that the multi-pitch detection performance and the drum transcription performance are on similar levels. When using 100ms as an onset tolerance, the transcription performance rises to 48% for multi-pitch detection and 47% for drum transcription. Given that current multi-pitch detection performance for pitched-only recordings of similar complexity is at about 60% (e.g. [17, 18]), we consider the transcription performance of the proposed system using complex pieces containing both pitched and unpitched elements very encouraging. Regarding the average precision and recall for multi-pitch detection, a similar trend is observed using the TRIOS recording, with a reported precision of 46.31% and a recall of 31.81%. It should be noted that for the drum components, the precision and recall are much more balanced.

Finally, we perform a comparison on multi-pitch detection performance using the method of [5], which is essentially the pitched component of the proposed model[1]. For the 'Take Five' recording, the method of [5] reached $F_{mp} = 75.68\%$ and for the RWC recordings the average F-measure is 36.64% (with 50ms tolerance), which indicates that modelling percussive instruments can actually increase the multi-pitch detection performance for music comprising both pitched and unpitched sound sources.

4. CONCLUSIONS

In this work, a novel system was proposed for the automatic transcription of polyphonic music containing both pitched and unpitched sounds. The system was able to detect multiple temporally overlapping notes as well as overlapping sounds from several drum kit components. We also created temporally aligned ground truth files for recordings from the RWC database that contain both harmonic and percussive sounds. Experiments on multi-pitch detection and drum transcription demonstrated encouraging results in both tasks, and also showed that the support of unpitched sounds can improve the multi-pitch detection performance for recordings containing both harmonic and percussive components. The source code for the proposed system is also available online[2].

In the future, we will further extend the proposed system and incorporate shift-invariance not only for harmonic sounds but also for percussive instruments; this way, the system will be enabled to account for tuning differences between drum kits. In addition, for further improving drum transcription performance, we will represent drum sounds as time-frequency patches and incorporate them into a joint model for pitched and unpitched music transcription. Finally, we will investigate the use of varying-Q time-frequency representations [19] for an improved temporal resolution of both high- and low-frequency content.

[1] It should be noted that the method of [5] had high scores and ranked first for the MIREX 2013 public transcription evaluations [17].

[2] https://code.soundsoftware.ac.uk/projects/pu_amt

5. REFERENCES

[1] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, appeared online July 2013.

[2] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence and Neuroscience*, 2008, Article ID 947438.

[3] G. Grindlay and D. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1159–1169, Oct. 2011.

[4] B. Fuentes, R. Badeau, and G. Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1854–1866, Sept. 2013.

[5] E. Benetos, S. Cherla, and T. Weyde, "An efficient shift-invariant model for polyphonic music transcription," in *6th International Workshop on Machine Learning and Music*, Prague, Czech Republic, Sept. 2013.

[6] H. Lindsay-Smith, S. McDonald, and M. Sandler, "Drumkit transcription via convolutive NMF," in *International Conference on Digital Audio Effects*, York, UK, Sept. 2012.

[7] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 529–540, 2008.

[8] J. Paulus and A. Klapuri, "Drum sound detection in polyphonicmusic with hidden Markov models," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009, Article ID 497292.

[9] J. Fritsch, "High quality musical audio source separation," M.S. thesis, UPMC / IRCAM / Telécom ParisTech, 2012.

[10] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: music genre database and musical instrument sound database," in *International Conference on Music Information Retrieval*, Baltimore, USA, Oct. 2003.

[11] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009, pp. 1869–1872.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[13] G. Grindlay and D. Ellis, "A probabilistic subspace model for multi-instrument polyphonic transcription," in *11th International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, Aug. 2010, pp. 21–26.

[14] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *11th International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, Aug. 2010, pp. 489–494.

[15] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[16] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conf.*, Barcelona, Spain, July 2010.

[17] "Music Information Retrieval Evaluation eXchange (MIREX)," `http://music-ir.org/mirexwiki/`.

[18] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a temporally-constrained shift-invariant model," *Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1727–1741, Mar. 2013.

[19] C. Schörkhuber, A. Klapuri, N Holighaus, and M. Dörfler, "A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *AES 53rd Conference on Semantic Audio*, London, UK, Jan. 2014.