City Research Online

City St George's, University of London

Citation: Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H. & Klapuri, A. (2012). Automatic Music Transcription: Breaking the Glass Ceiling. Paper presented at the 13th International Society for Music Information Retrieval Conference (ISMIR 2012), 8 - 12 Oct 2012, Porto, Portugal.

This is the unspecified version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/2093/

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

AUTOMATIC MUSIC TRANSCRIPTION: BREAKING THE GLASS CEILING

Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri †
Centre for Digital Music, Queen Mary University of London
{emmanouilb,simond,dimitrios,holger,anssik}@eecs.qmul.ac.uk

ABSTRACT

Automatic music transcription is considered by many to be the Holy Grail in the field of music signal analysis. However, the performance of transcription systems is still significantly below that of a human expert, and accuracies reported in recent years seem to have reached a limit, although the field is still very active. In this paper we analyse limitations of current methods and identify promising directions for future research. Current transcription methods use general purpose models which are unable to capture the rich diversity found in music signals. In order to overcome the limited performance of transcription systems, algorithms have to be tailored to specific use-cases. Semi-automatic approaches are another way of achieving a more reliable transcription. Also, the wealth of musical scores and corresponding audio data now available are a rich potential source of training data, via forced alignment of audio to scores, but large scale utilisation of such data has yet to be attempted. Other promising approaches include the integration of information across different methods and musical aspects.

1. INTRODUCTION

Automatic music transcription (AMT) is the process of converting an audio recording into some form of musical notation. AMT applications include automatic retrieval of musical information, interactive music systems, as well as musicological analysis [28]. Transcribing polyphonic music is a nontrivial task and while the problem of automatic pitch estimation for monophonic signals can be considered solved, the creation of an automated system able to transcribe polyphonic music without restrictions on the degree of polyphony or the instrument type still remains open. In this work we will be addressing the problem of polyphonic transcription; for an overview of melody transcription approaches the reader can refer to [39].

† Equally contributing authors. We acknowledge the support of the MIReS project, supported by the European Commission, FP7, ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711. E. Benetos is funded by a Queen Mary University of London Westfield Trust Research Studentship. D. Giannoulis and H. Kirchhoff are funded by a Queen Mary University of London CDTA Studentship.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2012 International Society for Music Information Retrieval.

The core problem for creating an AMT system is the detection of multiple concurrent pitches. In past years the majority of multi-pitch detection methods employed a combination of audio feature extraction and heuristic techniques, which also produced the best results in the MIREX multi-F0 (frame-wise) and note tracking evaluations [5, 33]. One commonly used technique of these methods is the iterative spectral subtraction approach of [27]. The best performing method in the MIREX multi-F0 and note tracking task is the work by Yeh [45], who proposed a joint pitch estimation algorithm based on a pitch candidate set score function, which is based on several audio features.

Another set of approaches formulates the frame-wise multiple-F0 estimation problem within a statistical framework. The problem can then be viewed as a maximum a posteriori (MAP) estimation problem:

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}\in\mathcal{C}} \mathcal{L}(\mathbf{c}|\mathbf{x}) \quad (1)$$

where $\mathbf{c} = \{F_0^1, \ldots, F_0^N\}$ is a set of fundamental frequencies, $\mathcal{C}$ is the set of all possible F0 mixtures, and $\mathbf{x}$ is the observed audio signal within a single analysis frame. If no prior information is specified, the problem can be expressed as a maximum likelihood (ML) estimation problem using Bayes' rule, e.g. [11, 14]. A related method was proposed in [37], using a generative model with a non-homogeneous Poisson process.

Finally, the majority of recent transcription papers utilise and expand *spectrogram factorisation* techniques (e.g. [7, 10]). Non-negative matrix factorisation (NMF) is a technique first introduced as a tool for music transcription in [43]. In its simplest form, the NMF model decomposes an input spectrogram $\mathbf{X} \in \mathbb{R}_+^{K\times N}$ with K frequency bins and N frames as:

$$\mathbf{X} = \mathbf{W}\mathbf{H} \quad (2)$$

where $\mathbf{W} \in \mathbb{R}_+^{K\times R}$ contains the spectral bases for each of the R pitches and $\mathbf{H} \in \mathbb{R}_+^{R\times N}$ is the pitch activity matrix across time. An alternative formulation of NMF called probabilistic latent component analysis (PLCA) has also been employed for transcription (e.g. [22]). In PLCA the matrices in the model are considered to be probability distributions, thus allowing for a model that can be easily extended and formalised. Additional transcription methods have been proposed in the literature, employing sparse coding techniques (e.g. [1]), genetic algorithms (e.g. [40]), and machine learning algorithms (e.g. [38]), which due to space limitations cannot be detailed here.

For note tracking, hidden Markov models (HMMs) are frequently used at a postprocessing stage (e.g. [38]). Other

Participants	2009	2010	2011
Yeh and Roebel	0.69	0.69	0.68
Dressler	-	-	0.63
Benetos and Dixon	-	0.47	0.57
Duan, Han, and Pardo	0.57	0.55	-

Table 1. Best results using the accuracy metric for the MIREX Multi-F0 estimation task, from 2009-2011. Details about the employed metric can be found in [33].

techniques include temporal smoothing (e.g. using a median filter) and minimum duration pruning [10].

In the remainder of this paper we analyse limitations of current approaches and identify promising directions for overcoming the obstacles in current performance.

2. CHALLENGES

Despite significant progress in AMT research, there exists no end-user application that can accurately and reliably transcribe music containing the range of instrument combinations and genres available in recorded music. The performance of even the most recent systems is still clearly below that of a human expert, who requires multiple takes, makes extensive use of prior knowledge and complex inference, and produces imperfect results. Furthermore, current test sets are limited in their complexity and coverage. Table 1 gives the results for the frame-based multiple-F0 estimation task of the MIREX evaluation [33]. Results for the note tracking task are much inferior, in the range of 0.2–0.35 average F-measure with onset-offset detection and 0.4–0.55 average F-measure with onset detection only. As we propose in Section 3, informing transcription via user-assistance or by providing a draft score in some applications are ways to increase systems' performance and overcome the observed plateau.

Currently proposed systems also fall short in flexibility to deal with diverse target data. Music genres like classical, heavy metal, hip-hop, ambient electronic and traditional Chinese music have little in common. Furthermore styles of notation vary with genre. For example Pop/Rock notation might represent melody, chords and (perhaps) bass line, whereas a classical score would usually contain all the notes to be played, and electroacoustic music has no standard means of notation. The task of tailoring AMT systems to specific styles has yet to be addressed. In Section 4 we propose systems focusing on instrument- or genre-specific transcription.

Algorithms are developed independently to carry out individual tasks such as multiple-F0 detection, beat tracking and instrument recognition. Although this is necessary, considering the complexity of each task, the challenge remains in combining the outputs of the algorithms, or better, combining the algorithms themselves to perform joint estimation of all parameters, in order to avoid the cascading of errors when the algorithms are combined sequentially. In Section 5, we propose the fusion of information across multiple musical aspects and the combination of methods targeting the same feature.

Another challenge concerns the availability of data for training and evaluation. Although there is no shortage of transcriptions and scores in standard music notation, human effort is required to digitise and time-align them to the recordings. Except for the case of solo piano, data sets currently employed for evaluation are small: a small subset from the RWC database [20] which contains only 12 tracks is commonly used (although the RWC database contains many more recordings) and the MIREX multi-F0 recording lasts only 54 seconds. Such small datasets cannot be considered representative; the danger of overfitting and thus overestimating system performance is high. It has been observed for several tasks that dataset developers tend to attain the best MIREX results [33]. In Section 6, we discuss ways to generate more training data.

At present, there is no established single unifying framework for music transcription as HMMs are for speech recognition. Likewise, there is no standard method for front end processing of the signal, with various approaches including STFT, constant Q transform [8] and auditory models, each leading to different mid-level representations. The challenge in this case is to characterise the impact of such design decisions on the AMT results. In Section 7, we consider the implications and steps required to progress from existing systems to complete transcription.

In addition to the above, the research community shares code and data on an ad hoc basis, with poor management and restrictive licensing limiting the level of re-use of research outputs. Many PhD students, for example, start from scratch spending valuable time "reinventing wheels" before proceeding to address current research issues. The lack of standard methodology is a contributing factor, with the multiplicity of approaches to AMT making it difficult to develop a useful shared code-base. The Reproducible Research movement [9], with its emphasis on open software and data, provides examples of best practice which are worthy of consideration by our community.

Finally, present research in AMT introduces certain challenges in itself that might constrain the evolution of the field. Advances in AMT research have mainly come from engineers and computer scientists, particularly those specialising in machine learning. Currently there is minimal contribution from computational musicologists, music psychologists or acousticians. Here the challenge is to integrate knowledge from these fields, either from the literature or by engaging these experts as collaborators in AMT research.

AMT research is quite active and vibrant at present, and we do not presume to predict what the state of the art will be in the next years and decades. In the remainder of the paper we propose promising techniques that can be utilised and further investigated in order to address the aforementioned limitations in transcription performance. In Fig. 1 we provide a general diagram of transcription, incorporating techniques discussed in the following sections.

3. INFORMED TRANSCRIPTION

3.1 Semi-automatic Approaches

Semi-automatic or *user-assisted transcription* refers to approaches where the user provides a certain amount of prior information to facilitate the transcription process [26]. Al-

though such systems are not applicable to the analysis of large music databases, they can be of use for musicians, musicologists, and—if a suitable synthesis method exists—for intelligent audio manipulation.

AMT systems usually have to solve a number of tasks, the nature of which depends on the type of music analysed and the level of detail required for the score representation. While some of these tasks might be quite easy for a human listener, it is often difficult to find an algorithmic formulation. The advantage of semi-automatic approaches is the fact that certain tasks that are inherently difficult to solve algorithmically can be assisted by the user of the system. Semi-automatic transcription systems might also pave the way for more robust fully-automatic ones, because the possibility of replacing the human part by an equally-performing computational solution always exists.

In principle any acoustic or score-related information that can facilitate the transcription process can act as prior information for the system. However, to be of use in a practical application, it is important that it does not require too much time and effort, and that the required information can be reliably extractable by the user, who might not be an expert musician.

Depending on the expertise of the targeted users, information that is easy to provide could include key, tempo and time signature of the piece, structural information, information about the instrument types in the recording, or even asking the user to label a number of notes for each instrument. Although many proposed transcription systems -often silently- make assumptions about certain parameters, such as the number or types of instruments in the recording, not many published systems explicitly incorporate prior information from a human user. In the context of source separation, Ozerov et al. [36] proposed a framework that enables the incorporation of prior knowledge about the number and types of sources, and the mixing model. The authors showed that by using prior information, a better separation can be achieved than with completely blind systems. A system for user-assisted music transcription was proposed in [26], where the user provides information about the instrument identities or labels a number of notes for each instrument. This knowledge enabled the authors to sidestep the error-prone task of source identification or timbre modelling, and to evaluate the proposed non-negative framework in isolation.

3.2 Score-informed Approaches

Contrary to speech, only a small fraction of music is fully spontaneous, as musical performances are typically based on an underlying composition or song. Although transcription is usually associated with the analysis of an unknown piece, there are certain applications for which a score is available, and in these cases the AMT system can exploit this additional knowledge [42]. For example in automatic instrument tutoring [6, 44], a system evaluates the performance of a student based on a reference score and provides feedback. Thus, the correctly played passages need to be identified, along with any mistakes made by the student, such as missed or extra played notes. Another example application is the analysis of expressive performance, where the tempo, dynamics, and timing deviations relative

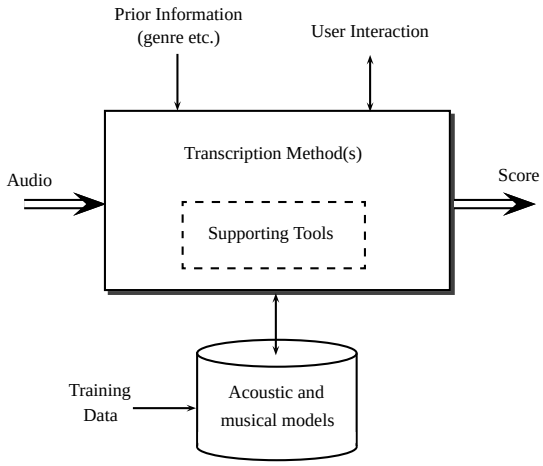

Figure 1. General overview of transcription. Supporting tools refer to techniques which can facilitate transcription, e.g. key estimation, instrument recognition.

to the score are the focus of the analysis. There are often small differences between the reference score and the performance, and in most cases, the score will not contain the absolute timing of notes and thus will need to be time-aligned with the recording as a first step.

One way to utilise the automatically-aligned score is for initialising the pitch activity matrix $\mathbf{H}$ in a spectrogram factorisation-based model (see Eq. (2)), and keeping these fixed while the spectral templates $\mathbf{W}$ are learnt, as in [16]. After the templates are learnt, the gain matrix can also be updated in order to cater for note differences between the score and the recording.

4. INSTRUMENT- AND GENRE-SPECIFIC TRANSCRIPTION

Current approaches for AMT usually employ instrument models that are not restricted to specific instrument types, but applicable and adaptable to a wide range of musical instruments. In fact, most transcription algorithms that are based on heuristic rules and those that employ human sound perception models even deliberately disregard specific timbral characteristics in order to enable an instrument-independent detection of notes. Even many so-called piano transcription methods are not so much tailored to piano music as *tested* on such music; they do not implement a piano-specific instrument model. Similarly, the aim of many transcription methods is to be applicable to a broad range of musical genres.

The fact that only a small number of publications on instrument- and genre-specific transcription exist, is particularly surprising when we compare AMT to the more mature discipline of automatic speech recognition. Continuous speech recognition systems are practically always language-specific and typically also domain-specific, and many modern speech recognisers include speaker adaptation [24].

Transcription systems usually try to model a wide range of musical instruments using a single set of computational methods, thereby assuming that those methods can be applied equally well to different kinds of musical instruments.

However, depending on the sound production mechanism of the instruments, their characteristics can differ considerably and might not be captured equally well by the same computational model or might at least require instrument-specific parameters and constraints if a common model is used. Furthermore, acoustic instruments incorporate a wide range of playing styles, which can differ notably in sound quality. On the other hand we can revert to the extensive literature on the physical modelling of musical instruments. A promising direction is to incorporate these models in the transcription process or at least use them as prior information that can then be adapted to the recording under analysis. Some examples of instrument-specific transcription are for violin [29], bells [30], tabla [19] and guitar [3].

The application of instrument-specific models, however, requires the target instrumentation either to be known or inferred from the recording. Instrument identification in a polyphonic context, as opposed to monophonic, is rendered difficult by the way the different sources blend with each other, with a high degree of overlap in the time-frequency domain. The task is closely related to sound source separation and as a result, many systems operate by first separating the signals of different instruments from the mixture or by generating time-frequency masks that indicate spectral regions that belong only to a particular instrument which can then be classified more accurately [13]. There are also systems that try to extract features directly from the mixture or by focusing on time-frequency regions with isolated note partials [4]. A review of instrument identification methods can be found in [34, sect. IV].

The advantage of restricting a transcription system to a certain musical genre lies in the fact that special (expert) knowledge about that genre can be incorporated. Musicological knowledge about structure (e.g. sonata form), harmony progressions (e.g. 12-bar blues) or specific instruments (e.g. Irish folk music) can enhance the transcription accuracy. Genre-specific AMT systems have been designed for genres such as Australian aborginal music [35]. In order to build a general-purpose AMT system, several genre-specific transcription systems could be combined and selected based on a preliminary genre classification stage.

5. INFORMATION INTEGRATION

5.1 Fusing information across the aspects of music

Many systems for note tracking combine multiple-F0 estimation with onset and offset detection, but disregard concurrent research on other aspects of music, such as instrumentation, rhythm, or tonality. These aspects are highly interdependent and they could be analysed jointly, combining information across time and across features to improve transcription performance.

A human transcriber interprets the performed notes in the context of a metrical structure consisting of a semi-regular, hierarchical system of accents. Extensive research has been performed into tempo induction, beat tracking and rhythm parsing [21], but transcription rarely takes advantage of this knowledge. An exception is the use of beat-synchronous features in chord transcription [31], where the audio is segmented according to the location of beats, and features are averaged over these beat-length intervals. The advantage of a more robust feature (less overlap between succeeding chords) is balanced by a loss in temporal resolution (harmonic change is assumed not to occur within a beat). For note transcription, it is unrealistic to assume that notes do not change within beats, but a promising approach would be to use a similar technique at a lower (i.e. sub-beat) metrical level, corresponding to the fastest note sequences. The resulting features would be more robust than frame-level features, and advantage could be taken of known (or learnt) rhythmic patterns and effects of metrical position.

Key is another high-level musical cue that, if known or estimated from the signal, provides useful prior information for the extraction of notes and chords. Key can be modelled as imposing a probability distribution over notes and chords for different metrical positions and durations. Therefore, by specifically modelling key, transcription accuracy can be improved, e.g. by giving more weight to notes which belong to the current key. Genre and style are also influential factors for modelling the distribution of pitch classes in a key. Several key estimation approaches have been proposed, but these are rarely exploited for AMT, with the exception of [41], which gave the best results for the MIREX 2008 note tracking task.

Likewise, local harmony (the current chord) can be used to inform note transcription. The converse problem, determining the chord given a set of detected notes, is also a transcription task. A chord transcription system which uses a probabilistic framework to jointly model the key, metre, chord and bass notes is presented in [31].

Finally, information can also be integrated over time. Most AMT systems to date have modeled only short-term dependencies, often using Markov models to describe expected melodic, harmonic and rhythmic sequences. As a notable exception, [32] utilized structural repetitions for chord transcription. Also the musical key establishes a longer-term (tonal) context for pitch analysis.

5.2 Combining methods targeting the same feature

Information could also be integrated by combining multiple estimators or detectors for a single feature, for instance combining two multi-pitch estimators, especially if these are based on different acoustic cues or different processing principles. This could help overcome weak points in the performance of the individual estimators, offer insight on the weaknesses of each and raise the overall system accuracy. In a different context, several pitched instrument onset detectors, which individually have high precision and low recall, have been successfully combined in order to obtain an improved detection accuracy [23]. For classification, adaptive boosting (AdaBoost) provides a powerful framework for fusing different classifiers in order to improve the performance [17].

5.3 Joint transcription and source separation

Source separation could be of benefit to transcription-related tasks such as instrument identification, where both tasks are interdependent, and accomplishing one of them could significantly ease the other. In this spirit, joint source separation and musical instrument identification methods have

been proposed using signal model-based probabilistic inference in the score-informed case [25]. Also, ideas and algorithms from the field of source separation can be utilised for AMT, especially regarding the exploitation of spatial information, if this is available [12, 36].

However, for most AMT tasks there is only one or two mixture signals available, and the number of sources is larger than the number of mixtures. In this case, the separation task is underdetermined, and can only be solved by requiring certain assumptions to hold for the sources. These could include sparsity, non-negativity and independence or they could involve structured spectral models like NMF models [22], spectral Gaussian scaled mixture models (Spectral-GSMMs) [2] or the source-filter model for sound production. Further constraints such as temporal continuity or harmonicity can be applied on spectral models. Techniques that employ spectral source modelling or an NMF-based framework that explicitly models the mixing process of the sources have been shown to perform well because they exploit the statistical diversity of the source spectrograms [2].

Finally, source separation can be fully utilised in a semi-supervised system like [12], where the user initially selects the desired audio source through the estimated F0 track of that source and subsequently the system refines the selected F0 tracks, and estimates and separates the relevant source.

6. CREATING TRAINING DATA

A large subset of AMT approaches perform experiments only on piano data, e.g. [10, 14, 38]. One reason is because it is relatively easy to create recordings with aligned ground-truth using e.g. a Disklavier. However, this emphasis on piano music sometimes leads to models that are tailored for pitched percussive instruments and could also be a cause for overfitting. Thus, ground-truth for multiple-instrument recordings is crucial for the further development of sophisticated transcription systems.

If musical scores become widely available in digital form (for example via crowd-sourced transcriptions), they provide valuable side-information for signal analysis, and in the extreme cases reduce the transcription task to the alignment of an existing score to the input audio, although it should be noted that different renditions of a song often vary considerably in their instrumentation and arrangement. One such example is the set of syncRWC annotations [1].

Most of the current AMT methods involve a training stage, where the parameters of the method are optimised using manually annotated data. The availability of recorded music with the exact underlying score opens up huge and largely unutilised opportunities for training complex models. In the case of genre- and instrument-specific transcription, separate parameter sets can be trained for different target material.

7. TOWARDS A COMPLETE TRANSCRIPTION

Most of the aforementioned transcription approaches tackle the problems of multiple-F0 estimation and note onset and offset detection. However, in order to fully solve the AMT problem and have a system that provides an output that is equivalent to sheet music, additional issues need to be addressed, such as metre induction, rhythm parsing, key finding, note spelling, dynamics, fingering, expression, articulation and typesetting. Although there are approaches that address many of these individual problems, there exists no 'complete' AMT system to date.

Regarding typesetting, current tools produce readable scores from MIDI data only (e.g. Lilypond [2]), however, cues from the music signal could also assist in incorporating additional information into the final score (e.g. expressive features for note phrasing). As far as dynamics are concerned, in [15] a method was proposed for estimating note intensities in a score-informed scenario. However, estimating note dynamics in an unsupervised way has not been tackled. Another issue would be the fact that most existing ground-truth does not include note intensities, which is difficult to annotate manually, except for datasets created using reproducing pianos (e.g. [38]), which automatically contain intensity information such as MIDI note velocities.

Recent work [3] addresses the problem of automatically extracting the fingering configurations for guitar recordings in an AMT framework. For computing fingering, information from the transcribed signal as well as instrument-specific knowledge is needed. Thus, a robust instrument identification system would need to be incorporated for computing fingerings in multi-instrument recordings.

For extracting expressive features, some work has been done in the past, mostly in the score-informed case. In [18] a framework for extracting expressive features both from a score-informed and an uninformed perspective is proposed. For the latter, an AMT system is used prior to the extraction of expressive features. It should be mentioned though that the extracted features (e.g. auditory loudness, attack) do not necessarily correspond to expressive notation. Thus, additional work needs to be done in order to provide a mapping between mid-level features and actual expressive markings in a transcribed music score.

8. CONCLUSIONS

Automatic music transcription is a rapidly developing research area where several different approaches are still being actively investigated. However from the perspective of evaluation results, the performance seems to converge towards a level that is not satisfactory for all applications.

One viable way of breaking the glass ceiling is to insert more information into the problem. For example, genre- or instrument-specific transcription allows the utilisation of high-level models that are more precise and powerful than their more general counterparts. A promising research direction is to combine several processing principles, or to extract various types of musical information, such as the key, metrical structure, and instrument identities, and feed that into a model that provides context for the note detection process. To enable work in this area, sharing code and data between researchers becomes increasingly important.

Note detection accuracy is not the only determining factor that enables meaningful end-user applications. Often

[1] http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/SyncRWC/

[2] http://lilypond.org/

it is possible to circumvent the limitations of the underlying technology in creative ways. For example in semi-automatic transcription, the problem is redefined as achieving the required transcription accuracy with minimal user effort. It is important to have end-user applications that drive the development of AMT technology and provide it with relevant feedback.

9. REFERENCES

[1] S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by non-negative sparse coding of power spectra. In *ISMIR*, pages 318–325, 2004.

[2] S. Arberet, A. Ozerov, F. Bimbot, and R. Gribonval. A tractable framework for estimating and combining spectral source models for audio source separation. *Signal Processing*, 92(8):1886–1901, 2012.

[3] A.M. Barbancho, A. Klapuri, L.J. Tardon, and I. Barbancho. Automatic transcription of guitar chords and fingering from audio. *IEEE TASLP*, 20(3):915–921, 2012.

[4] J.G.A. Barbedo and G. Tzanetakis. Musical instrument classification using individual partials. *IEEE TASLP*, 19(1):111–122, 2011.

[5] M. Bay, A. F. Ehmann, and J. S. Downie. Evaluation of multiple-F0 estimation and tracking systems. In *ISMIR*, pages 315–320, 2009.

[6] E. Benetos, A. Klapuri, and S. Dixon. Score-informed transcription for automatic piano tutoring. In *EUSIPCO*, 2012.

[7] N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE TASLP*, 18(3):538–549, 2010.

[8] J.C. Brown. Calculation of a constant Q spectral transform. *JASA*, 89(1):425–434, 1991.

[9] J. B. Buckheit and D. L. Donoho. WaveLab and reproducible research. Technical Report 474, Dept of Statistics, Stanford Univ., 1995.

[10] A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *ISMIR*, pages 489–494, 2010.

[11] Z. Duan, B. Pardo, and C. Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE TASLP*, 18(8):2121–2133, 2010.

[12] J.L. Durrieu and J.P. Thiran. Musical audio source separation based on user-selected F0 track. In *LVA/ICA*, pages 438–445, 2012.

[13] J. Eggink and G.J. Brown. A missing feature approach to instrument identification in polyphonic music. In *ICASSP*, volume 5, pages 553–556, 2003.

[14] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE TASLP*, 18(6):1643–1654, 2010.

[15] S. Ewert and M. Müller. Estimating note intensities in music recordings. In *ICASSP*, pages 385–388, 2011.

[16] S. Ewert and M. Müller. Using score-informed constraints for NMF-based source separation. In *ICASSP*, pages 129–132, 2012.

[17] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *JSAI*, 14(771-780):1612, 1999.

[18] R. Gang, G. Bocko, J. Lundberg, S. Roessner, D. Headlam, and M.F. Bocko. A real-time signal processing framework of musical expressive feature extraction using MATLAB. In *ISMIR*, pages 115–120, 2011.

[19] O. Gillet and G. Richard. Automatic labelling of tabla signals. In *ISMIR*, 2003.

[20] M. Goto. Development of the RWC music database. In *18th Int. Congress Acoustics*, pages 553–556, 2004.

[21] F. Gouyon and S. Dixon. A review of automatic rhythm description systems. *CMJ*, 29(1):34–54, 2005.

[22] G. Grindlay and D. Ellis. Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments. *IEEE JSTSP*, 5(6):1159–1169, 2011.

[23] A. Holzapfel, Y. Stylianou, A.C. Gedik, and B. Bozkurt. Three dimensions of pitched instrument onset detection. *IEEE TASLP*, 18(6):1517–1527, 2010.

[24] X. Huang, A. Acero, and H.-W. Hon, editors. *Spoken Language Processing: A guide to theory, algorithm and system development*. Prentice Hall, 2001.

[25] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H.G. Okuno. Simultaneous processing of sound source separation and musical instrument identification using Bayesian spectral modeling. In *ICASSP*, pages 3816–3819, 2011.

[26] H. Kirchhoff, S. Dixon, and A. Klapuri. Shift-variant non-negative matrix deconvolution for music transcription. In *ICASSP*, 2012.

[27] A. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE TASLP*, 11(6):804–816, 2003.

[28] A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription*. Springer, 2006.

[29] A. Loscos, Y. Wang, and W.J.J. Boo. Low level descriptors for automatic violin transcription. In *ISMIR*, pages 164–167, 2006.

[30] M. Marolt. Automatic transcription of bell chiming recordings. *IEEE TASLP*, 20(3):844–853, 2012.

[31] M. Mauch and S. Dixon. Simultaneous estimation of chords and musical context from audio. *IEEE TASLP*, 18(6):1280–1289, 2010.

[32] M. Mauch, K. Noland, and S. Dixon. Using musical structure to enhance automatic chord transcription. In *ISMIR*, pages 231–236, 2009.

[33] Music Information Retrieval Evaluation eXchange (MIREX). http://music-ir.org/mirexwiki/, 2011.

[34] M. Müller, D. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE JSTSP*, 5(6):1088–1110, 2011.

[35] A. Nesbit, L. Hollenberg, and A. Senyard. Towards automatic transcription of Australian aboriginal music. In *ISMIR*, pages 326–330, 2004.

[36] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE TASLP*, 20(4):1118–1133, 2012.

[37] P.H. Peeling and S.J. Godsill. Multiple pitch estimation using non-homogeneous Poisson processes. *IEEE JSTSP*, 5(6):1133–1143, 2011.

[38] G. Poliner and D. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP JASP*, 8:154–162, 2007.

[39] G. Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE TASLP*, 15(4):1247–1256, 2007.

[40] G. Reis, N. Fonseca, F. F. de Vega, and A. Ferreira. Hybrid genetic algorithm based on gene fragment competition for polyphonic music transcription. In *Conf. Applications of Evolutionary Computing*, pages 305–314. 2008.

[41] M.P. Ryynänen and A. Klapuri. Polyphonic music transcription using note event modeling. In *WASPAA*, pages 319–322, 2005.

[42] E.D. Scheirer. Using musical knowledge to extract expressive performance information from audio recordings. In H. Okuno and D. Rosenthal, editors, *Readings in Computational Auditory Scene Analysis*. Lawrence Erlbaum, 1997.

[43] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *WASPAA*, pages 177–180, 2003.

[44] Y. Wang and B. Zhang. Application-specific music transcription for tutoring. *IEEE MultiMedia*, 15(3):70–74, 2008.

[45] C. Yeh. *Multiple fundamental frequency estimation of polyphonic recordings*. PhD thesis, Université Paris VI - Pierre et Marie Curie, France, 2008.