MISSING TEMPLATE ESTIMATION FOR USER-ASSISTED MUSIC TRANSCRIPTION

Holger Kirchhoff [*1] *Simon Dixon*[1] *Anssi Klapuri*[2,3]

[1] Centre for Digital Music, Queen Mary University of London, UK
[2] Ovelin, Helsinki, Finland
[3] Tampere University of Technology, Finland

ABSTRACT

For a user-assisted music transcription system in which the user is asked to label some notes for each instrument in the recording, we investigate ways to limit the amount of information the user has to provide. Different methods are proposed and experimentally compared that enable the estimation of template spectra at pitch positions that have not been annotated by the user, in order to derive a full set of instrument templates that can be used within a non-negative matrix factorisation framework. A set of error metrics is presented that enables the evaluation of the NMF gain matrix. The results show that purely data-driven methods outperform more refined instrument models when the user annotates notes at many different pitches for each instrument. When notes are labelled at a smaller number of different pitches, the highest accuracies are obtained using pre-stored instrument templates that are adapted to the instruments in the mixture.

Index Terms— user-assisted music transcription, template estimation, source-filter model, adapting instrument spectra

1. INTRODUCTION

Automatic music transcription describes the computational process of transforming a recording of a piece of music into some form of symbolic notation. While transcription algorithms for monophonic instrument recordings achieve satisfactory results, the automatic transcription of polyphonic music remains an open challenge [1].

Semi-automatic or *user-assisted music transcription* refers to systems in which the user is actively involved in the transcription process by providing prior information about the underlying mixture under analysis or by interacting with intermediate transcription results. The types of information a musically-trained user can provide about a target recording are manifold and can include musical aspects such as key, tempo, time signature or structural information. Of particular benefit is information that facilitates the creation of accurate timbre models that in turn enable the identification of note objects of the underlying instruments in the recording.

In this paper we consider the case that a user has labelled a few notes for each instrument in the mixture which can be used to infer timbre models for the instruments. In a practical application this can be facilitated by presenting users with a piano-roll representation obtained by a fully-automatic transcription system and asking them to assign a few notes to each instrument. By means of a non-negative matrix factorisation framework this information can be utilised to extract prototype spectra for each labelled pitch of each instrument.

[*]This work was funded by a Queen Mary University of London CDTA studentship.

In order for a semi-automatic transcription system to be useful in practice, it is important to limit the amount of information the user has to provide. This means that a user cannot be expected to label notes at each pitch of each instrument occurring in the mixture. Spectral templates at pitches for which a user has not labelled any notes — here referred to as *missing templates* — need to be estimated from the existing spectral templates.

In prior work on user-guided music transcription [2], a system was proposed that performs automatic melody, bass, chord and drum transcription and then provides the user with various manipulation options. The user input was employed to modify the initial results without taking further evidence from the audio data into account. User-assisted techniques have predominantly been applied to the related field of audio source separation. Smaragdis and Mysore [3] used a hummed melody from the user to separate the corresponding instrument voice from the mixture. Barry et al. [4] separated instrument tracks from a stereo recording based on a user-specified azimuth range in the stereo panorama. Ozerov et al. [5] required information about the number of components per source and a source activity segmentation from the user. Fuentes et al. [6] asked the user to select the notes to be separated from an initial automatic transcription result. Likewise, Durrieu and Thiran [7] enable the user to select and modify pitch contours of the main melody to be separated. Approaches in [6, 7] only separated notes and contours specified by the user. No generalisation to unlabelled notes was performed, making the separation of longer excerpts laborious.

The remainder of this paper is structured as follows: In the following section, different methods for the estimation of missing spectral templates are presented. In Sect. 3, the evaluation procedure, the employed test sets and the metrics are explained and results are discussed. Conclusions are drawn in Sect. 4.

2. ESTIMATION METHODS

In this section, we discuss several methods for the estimation of missing templates, under the assumption that a few typical spectra at various pitches are available for each instrument, and spectra between these pitches need to be estimated. For this paper, we employ the non-negative framework from [8]. In this framework, each instrument is represented by a set of basis functions corresponding to different fundamental frequencies in the constant-Q spectrogram:

$$\mathbf{V} \approx \mathbf{\Lambda} = \sum_{i=1}^{I} \sum_{\phi=0}^{\Phi-1} \mathbf{w}_{i,\phi}^{\phi\downarrow} \, \mathbf{h}_{i,\phi}^{\top}. \tag{1}$$

In this equation, $\mathbf{V} \in \mathcal{R}_{+}^{K \times N}$ denotes the constant-Q magnitude spectrogram and $\mathbf{\Lambda} \in \mathcal{R}_{+}^{K \times N}$ the approximation by the model. $\mathbf{w}_{i,\phi} \in \mathcal{R}_{+}^{K}$ is a column vector containing the spectral template

of instrument i for a particular pitch ϕ. In this paper, a sub-semitone pitch resolution of 4 pitches per semitone is employed. Due to the logarithmic frequency axis, the distances between adjacent harmonic partials are independent of the fundamental frequency. We align all templates $\mathbf{w}_{i,\phi}$, so that the first partial is located at the first vector index and likewise all other partials appear at the same vector indexes regardless of the pitch, which has advantages for several estimation methods described in Sect. 2. The operator $\phi\downarrow$ denotes a downward shift (upward in pitch) of the vector elements by ϕ rows while the first ϕ rows are filled with zeros. $\mathbf{h}_{i,\phi} \in \mathcal{R}_+^N$ contains the activations (gains) of each basis function for instrument i at pitch ϕ. For convenience, we combine all vectors of an instrument i into a matrix $\mathbf{W}^i \in \mathcal{R}_+^{K\times\Phi}$, where $\mathbf{W}^i = [\mathbf{w}_{i,0}, \mathbf{w}_{i,1}, \ldots, \mathbf{w}_{i,\Phi-1}]$. Likewise, the gain matrix of an instrument i is here expressed by $\mathbf{H}^i \in \mathcal{R}_+^{\Phi\times N}$, where $\mathbf{H}^i = [\mathbf{h}_{i,0}, \mathbf{h}_{i,1}, \ldots, \mathbf{h}_{i,\Phi-1}]^\top$.

2.1. Copying

A very simple yet effective method to derive spectral templates at missing pitches is to employ a translated version of the user-provided spectrum at the nearest pitch. Within matrix $\mathbf{W}^i$, where all spectra are aligned, this can be achieved by merely copying the spectra to their adjacent pitch positions. This method assumes that the partial amplitudes of near pitches are approximately the same, which is also the underlying principle of the shift-invariant NMF algorithm [9].

2.2. Interpolation

Another data-driven approach is to estimate the missing spectra by interpolating between the existing spectra. We examine two different interpolation methods. The easiest way to interpolate missing spectral templates is to apply linear interpolation to each spectral bin of the aligned spectra in matrix $\mathbf{W}^i$. The interpolation can thus be applied along the pitch axis to each frequency bin separately. We call this *plain interpolation*.

Another interpolation method takes several spectra in the pitch vicinity of the missing spectrum into account. For each missing template, a weighted average of surrounding templates is computed using a Hann window centred at the missing template for the weights. We empirically chose a window length of 9 semitones. When no provided spectrum falls within the hanning window range, the missing spectra are estimated by copying (Sect. 2.1).

2.3. Source-filter model

The source-filter model was originally introduced for speech synthesis [10] but it has also been used extensively for musical instrument modelling both for analysis and synthesis purposes (e. g. [11, 12, 13, 14]). A good introduction can be found in [15].

The source-filter model assumes that the sound production process of an acoustic source consists of two distinct parts: A *generator* or *source* that produces an excitation signal, and a *resonator* or *filter* that shapes the excitation signal. This assumption does not hold equally well for all types of musical instruments. It is a good fit for bowed string instruments (i. e. the violin family), in which source and filter — the vibrating string and the instrument body — are reasonably well decoupled. It holds less for instruments with stronger interdependencies between the source and the filter part — such as many woodwind and brass instruments. Nevertheless, the source-filter model is able to capture characteristics of the instruments that can either be modelled as a function of partial index (e. g. weak even harmonics in clarinet spectra) or absolute frequency (e. g. formants and resonances of the instrument body).

Recently, the source-filter model has been integrated into the NMF framework in various ways (e. g. [11, 13, 16]) to reduce the number of parameters to estimate. Here we propose an implementation[1] that estimates the model parameters based on the β-divergence between the original and the modelled spectra and that operates on isolated instrument spectra as opposed to being integrated in an NMF framework. It is inspired by the methods in [11] and [12].

The source-filter model proposed here approximates the (source) excitation spectrum $\mathbf{e}$ and the filter spectrum $\mathbf{h}$ from a number of provided instrument spectra $\mathbf{w}_p$ ($p \in [1, \ldots, P]$) at different pitches ϕ_p according to the following equation:

$$\mathbf{w}_p \approx \hat{\mathbf{w}}_p = s_p \cdot {}^{\phi_p\downarrow}\mathbf{e} \otimes \mathbf{h}. \tag{2}$$

In this equation, the $\otimes$ operator denotes elementwise multiplication of the vectors. s_p is a scaling factor that compensates for gain differences among the provided instrument spectra. The pitch ϕ_p of each spectrum $\mathbf{w}_p$ is here expressed in terms of frequency bin indices of the fundamental frequency. The operator $\phi_p\downarrow$ translates the excitation spectrum along the logarithmic frequency axis to the correct pitch position ϕ_p as described above. The scaling factors s_p for all pitches can be combined into a single vector $\mathbf{s}$ of length P.

We combine all vectors $\mathbf{w}_p$ from which $\mathbf{s}$, $\mathbf{e}$ and $\mathbf{h}$ are estimated into a matrix $\mathbf{W}' \in \mathcal{R}_+^{K,P}$. Likewise, $\hat{\mathbf{W}}'$ denotes a matrix with the same dimensions that contains in its columns all estimated templates $\hat{\mathbf{w}}_p$ based on the current estimates of $\mathbf{s}$, $\mathbf{e}$ and $\mathbf{h}$ according to Eq. 2.

Based on the provided instrument spectra at distinct pitches, the model estimates the three vectors $\mathbf{s}$, $\mathbf{e}$ and $\mathbf{h}$. This is achieved by randomly initialising the three vectors and iteratively applying gradient descent on each vector. The β-divergence is used as cost function. For the evaluation in Sect. 3, we set $\beta = 0$.

The update equations for the individual components of $\mathbf{s}$, $\mathbf{e}$ and $\mathbf{h}$ are given by

$$s_p \leftarrow s_p \cdot \frac{\sum_{k=1}^{K} {W'}_{k,p} {\hat{W}'}_{k,p}^{\beta-2} e_{k-\phi_p} h_k}{\sum_{k=1}^{K} {\hat{W}'}_{k,p}^{\beta-1} e_{k-\phi_p} h_k} \tag{3}$$

$$e_k \leftarrow e_k \cdot \frac{\sum_{p=1}^{P} {W'}_{k+\phi_p,p} \cdot {\hat{W}'}_{k+\phi_p,p}^{\beta-2} \cdot s_p \cdot h_{k+\phi_p}}{\sum_{p=1}^{P} {\hat{W}'}_{k+\phi_p,p}^{\beta-1} \cdot s_p \cdot h_{k+\phi_p}} \tag{4}$$

$$h_k \leftarrow h_k \cdot \frac{\sum_{p=1}^{P} {W'}_{k,p} \cdot {\hat{W}'}_{k,p}^{\beta-2} \cdot s_p \cdot e_{k-\phi_p}}{\sum_{p=1}^{P} {\hat{W}'}_{k,p}^{\beta-1} \cdot s_p \cdot e_{k-\phi_p}} \tag{5}$$

A detailed derivation of these update equations can be found in [17].

Note that the model in Eq. 2 contains two ambiguities which need to be addressed in order to provide unique results for $\mathbf{s}$, $\mathbf{e}$ and $\mathbf{h}$: First, scaling $\mathbf{e}$ by a constant factor and either $\mathbf{s}$ or $\mathbf{h}$ by the inverse of this factor results in same estimates of the spectra. And second, multiplying one of the vectors by an exponential function and dividing the other by the same function likewise yields the same estimated spectra. Details about these ambiguities and ways to fix them can be found in [17].

The filter response $\mathbf{h}$ can only be reliably estimated at those frequency positions where the spectral energy is sufficiently greater than the noise floor, that is at the harmonic partial frequencies of the provided spectra. In order to obtain a *continuous* filter curve, a smoothed interpolation is applied to the estimated amplitudes. A

[1] available from: `http://code.soundsoftware.ac.uk/projects/sourcefiltermodel`

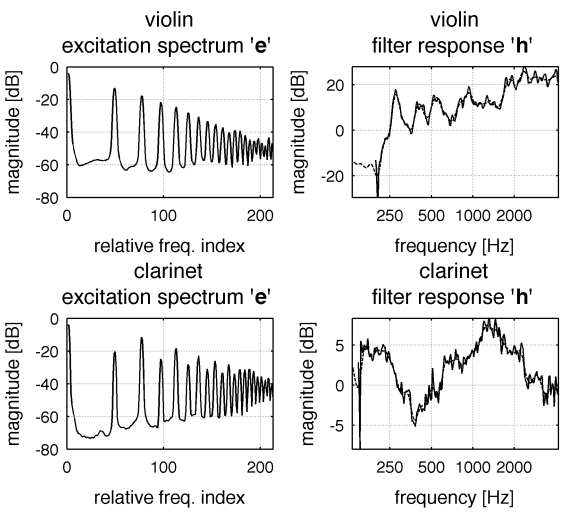

Fig. 1. Source-filter model example for violin and clarinet. The excitation spectra $\mathbf{e}$ are shown on the left, the original (solid line) and smoothed (dashed line) filter responses $\mathbf{h}$ are displayed on the right.

number of cosine functions are fitted to the amplitudes and a regularisation parameter is utilised that prevents the cosine approximation to take on steep slopes. This technique was introduced in [18] as *discrete cepstrum spectral envelope*. In our implementation 40 cosine functions and a regularisation parameter of $5 \cdot 10^{-4}$ were employed.

Figure 1 displays the results of the estimation of the excitation signal $\mathbf{e}$ and the filter response $\mathbf{h}$ for two different instruments from the RWC database [19].

2.4. Adapting database templates

In [8] we showed that considerably higher transcription accuracies can be achieved if spectral templates are learned directly from the recording under analysis as opposed to a database of instruments. However, database templates might be useful for the estimation of spectra at missing pitches, as they can provide evidence about typical spectra of the instruments without employing an explicit instrument model. We assume here that the differences between recordings of instruments of the same type are either caused by varying recording conditions or by differences in instrument construction and that these differences can be summarised by a single linear time-invariant system. Given a few spectral templates extracted from the recording, a single filter can be estimated that adapts the database templates to the extracted pitches. This filter can then also be applied to adapt templates at all other pitches[2].

In mathematical form, the adaptation of database spectra to the spectra of the recording can be expressed by

$$\mathbf{w}_{\text{data},p} \approx \hat{\mathbf{w}}_{\text{data},p} = \mathbf{w}_{\text{DB},p} \otimes \mathbf{f}. \tag{6}$$

In this equation, $\mathbf{w}_{\text{data},p}$ denotes the spectra estimated from the recording at pitches ϕ_p with $p \in [1, \ldots, P]$. $\hat{\mathbf{w}}_{\text{data},p}$ is the approximation of these spectra resulting from the elementwise multiplication of the database spectra $\mathbf{w}_{\text{DB},p}$ with the filter response $\mathbf{f}$.

[2]An implementation is available from: `http://code.soundsoftware.ac.uk/projects/adaptinstrspec`

The aim of the estimation procedure is to determine $\mathbf{f}$ in such a way that the error between the original spectra and the filtered database spectra is minimised. Here, we again apply the β-divergence cost function which generalises various well known cost functions. We use $\beta = 0$ for the evaluation in Sect. 3. Gradient descent is applied to estimate the filter response and the update equation is given by (cf. [20]):

$$\mathbf{f} \leftarrow \mathbf{f} \otimes \frac{\sum_{p=1}^{P} \mathbf{w}_{\text{data},p} \otimes \hat{\mathbf{w}}_{\text{data},p}^{\beta-2} \otimes \mathbf{w}_{\text{DB},p}}{\sum_{p=1}^{P} \hat{\mathbf{w}}_{\text{data},p}^{\beta-1} \otimes \mathbf{w}_{\text{DB},p}}. \tag{7}$$

For the estimation of $\mathbf{f}$, only the peak amplitudes of the partials are considered. Here, the same problem arises as in the case of the source filter model in Sect. 2.3: if the number of spectra $\mathbf{w}_{\text{data},p}$ is small, $\mathbf{f}$ will not be estimated at all frequency bins k and the filter response needs to be interpolated. The same cosine approximation of the filter reponse as in Sect. 2.3 is therefore applied to interpolate and smooth the filter response. In this case, 20 cosine coefficients are used and a regularisation parameter of 0.001 is applied to allow for larger amplitude variations than in Sect. 2.3.

3. EVALUATION

3.1. Procedure

The missing spectrum estimation methods described above were experimentally compared to determine how accurately each set of estimated spectra represent the actual instruments, using a transcription context where only a few spectra can be extracted from the user information. To that end, spectral templates were extracted based on the pitch information about *all* notes of *all* instruments in each recording. To simulate user-labelled notes at different pitch resolutions, a certain number of spectra were systematically discarded for each instrument: only every second, third, fourth, etc. instrument spectrum was preserved for each instrument. All estimation methods were then applied to these reduced sets of instrument spectra in order to obtain template estimates for all pitches of each instrument, and these complete sets were used to obtain the gain matrices $\mathbf{H}^i$ by making use of the above mentioned non-negative framework. This procedure was applied to several datasets and with various amounts of discarded spectra.

3.2. Datasets

Three datasets were used for the evaluation: The MIREX development set for the *multiple-f0 estimation and tracking* task [21], the Bach10 dataset [22] and the Trios dataset [23]. The MIREX dataset consists of a single 54 s recording of a Beethoven string quartet arranged for a woodwind quintet, the Bach10 dataset [22] contains ten 4-part Bach chorales with lengths between 25 s and 41 s, played by four different instruments, and the Trios dataset [23] consists of recordings of pieces by Mozart, Schubert, Brahms, Lussier and a jazz piece by Paul Desmond, all played by three instruments with various instrumentations. The jazz piece from the Trios dataset was omitted as one of the instruments is a drum set which is not considered in our analyses. All datasets are available online.

3.3. Metrics

Common transcription measures assume hard decisions on the active pitches per frame which require specific thresholding or tracking techniques. In this paper, however, we are interested in the effect of

different template dictionaries on the NMF gain matrices. Any measure that compares detected pitches to ground-truth pitches per frame also depends on the applied decision-making process. Results lose their generality if there is no one-to-one correspondence between the quality of the NMF gain matrix and the detected pitches.

We propose here a set of error measures that is capable of evaluating the quality of a gain matrix as well as being able to deal with parts-based transcriptions where the transcribed notes are assigned to their respective instruments. In our case and also more generally for NMF algorithms, gain matrices can be seen as pitch detection functions since they aim at containing high values at time-frequency positions where notes are present and low values elsewhere.

The *gain precision* computes for each instrument i the amount of energy in the gain matrix $\mathbf{H}^i$ that is concentrated in the ground truth fundamental frequencies and relates it to the overall energy in the matrix:

$$\mathrm{GP}_i = \frac{\sum_{n=1}^{N} \sum_{\phi \in \mathcal{F}_{n,i}} \left([\mathbf{H}^i]_{\phi,n} \right)^2}{\sum_{n=1}^{N} \sum_{\phi'=1}^{\Phi} \left([\mathbf{H}^i]_{\phi',n} \right)^2}. \quad (8)$$

In this equation, $\mathcal{F}_{n,i}$ denotes the set of annotated ground truth pitches in the n-th frame for instrument i.

Similar to the gain precision, the *gain recall* measures the amount of energy assigned to right instruments:

$$\mathrm{GR}_i = \frac{\sum_{n=1}^{N} \sum_{\phi \in \mathcal{F}_{n,i}} \left([\mathbf{H}^i]_{\phi,n} \right)^2}{\sum_{n=1}^{N} \sum_{\phi \in \mathcal{F}_{n,i}} \sum_{i'=1}^{I} \left([\mathbf{H}^{i'}]_{\phi,n} \right)^2}. \quad (9)$$

I here denotes the total number of instruments in the mixture.

The *gain f-measure* combines the two measures described above. It is given by

$$\mathrm{GF}_i = 2 \cdot \frac{\mathrm{GP}_i \cdot \mathrm{GR}_i}{\mathrm{GP}_i + \mathrm{GR}_i}, \quad (10)$$

and ranges between one and zero. Ideally, we would like to see all energy concentrated in the fundamental frequencies and assigned to the right instrument, which corresponds to $\mathrm{GF}_i = 1$.

3.4. Results

Figure 2 shows the results of the different estimation methods. The left column displays the results when spectral templates at *all* pitches are present, whereas the right column shows results for the case in which the user has only provided every 7th pitch. Results for other pitch resolutions are not displayed here due to space limitations.

In the case where notes are labelled at a high pitch resolution (0 skipped pitches), the purely data-driven methods *copying* and *interpolation* obtain the highest gain f-measures. Note that spectra are estimated on a sub-semitone pitch resolution in order to accommodate tuning differences and capture pitch contours. Therefore templates between the notes need to be estimated even when all notes are labelled. All the remaining estimation methods can even modify the provided spectra in different ways, which has an immediate effect on the per-instrument transcription accuracies. The same trend was observed when 1 or 2 pitches were skipped (not displayed here).

When it comes to lower pitch resolutions, the purely data driven methods produce less accurate results and are outperformed by the *adapted database spectra*. In general, results of the method of *adapting database spectra* do not vary much when spectra are provided at different pitch resolutions. For this method, the pitch resolution only determines the number of spectra that are available for estimating the response $\mathbf{f}$ in Eq. 6.

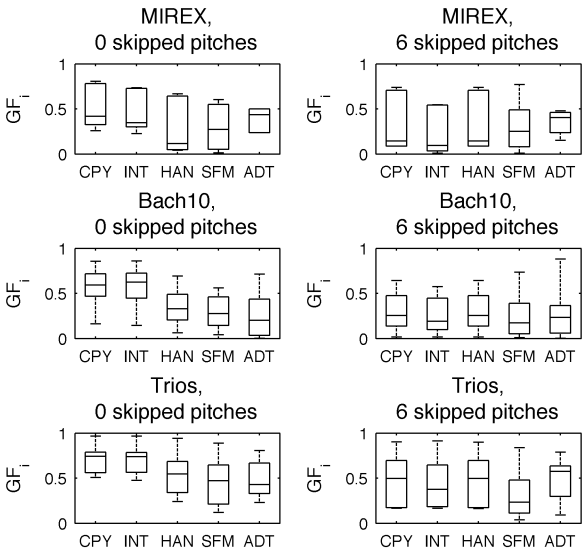

Fig. 2. Results for the different missing spectrum estimation methods for 3 different datasets based on spectral templates at different pitch resolutions. On the left-hand side, notes at all pitches were provided by the user; on the right-hand side, only spectra at every 7th pitch were provided. In each panel the different estimation methods can be compared: copying (CPY), plain interpolation (INT), hanning interpolation (HAN), source-filter model (SFM) and adapting database templates (ADT).

For all pitch resolutions, the *source-filter model* produces the lowest transcription accuracies. The proportion of woodwind instruments — for which the source-filter model is only a coarse fit — was comparably large in the employed datasets which might explain this loss in accuracy. In addition, at low pitch resolutions the small number of available spectra does not allow for an accurate estimation of the source and filter parts.

4. CONCLUSION

For the task of user-assisted music transcription, we investigated the use case in which the user labels a few notes for every instrument in the recording. More precisely we looked at ways to minimise the amount of information the user has to provide — particularly the pitch resolution at which note labels should be provided. Different methods for the estimation of template spectra at pitch positions that have not been provided by the user were experimentally compared: data-driven methods, such as copying existing spectra to adjacent pitch positions or interpolating partial amplitudes have been compared to more refined instrument models, such as the source-filter model and an adaptation of pre-learned spectra of the same instrument type. For the evaluation, a set of error measures for NMF gain matrices was proposed that could equally well be applied to any pitch detection function of a parts-based transcription. The methods were experimentally compared on three different datasets and with varying pitch resolutions. The results suggest that the data-driven methods *copying* and *interpolation* work well when instrument templates are available at a higher pitch resolution where each template only needs to be used within a comparably small pitch range. At lower pitch resolutions it seems to be better to revert to previously learned database templates and adapt them to the specific instruments in the mixture.

5. REFERENCES

[1] Anssi Klapuri and Manuel Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, 2006.

[2] C. Dittmar and J. Abeßer, "Automatic music transcription with user interaction," in *34. Deutsche Jahrestagung für Akustik (DAGA)*, 2008, pp. 567–568.

[3] P. Smaragdis and G. J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, October 2009.

[4] D. Barry, B. Lawlor, and E. Coyle, "Real-time sound source separation: Azimuth discrimination and resynthesis," in *117th Audio Engineering Society Convention*, San Francisco, October 2004.

[5] A. Ozerov, C. Févotte, R. Blouet, and J.L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, Prague, May 2011.

[6] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 2654—2658.

[7] J.L. Durrieu and J.P. Thiran, "Musical audio source separation based on user-selected f0 track," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel-Aviv, Israel, March 2012.

[8] H. Kirchhoff, S. Dixon, and A. Klapuri, "Shift-variant non-negative matrix deconvolution for music transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012.

[9] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted non-negative matrix factorisation for sound source separation," in *IEEE Workshop on Statistical Signal Processing*, 2005, pp. 1132–1137.

[10] H. Dudley, "Remaking speech," *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[11] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*, 2006.

[12] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[13] Romain Hennequin, Roland Badeau, and Bertrand David, "NMF with time-frequency activations to model nonstationary audio events," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744–753, May 2011.

[14] Marcelo Caetano and Xavier Rodet, "A source-filter model for musical instrument sound transformation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2012, pp. 137–140.

[15] M. Müller, D.P.W. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, October 2011.

[16] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[17] H. Kirchhoff, S. Dixon, and A. Klapuri, "Derivation of update equations for a source-filter model based on beta-divergence," Tech. Rep. C4DM-TR-10-12, Queen Mary University of London, 2012, Available from: `http://www.eecs.qmul.ac.uk/~holger/C4DM-TR-10-12`.

[18] Diemo Schwarz, "Spectral envelopes in sound analysis and synthesis," M.S. thesis, Universität Stuttgart, 1998.

[19] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, pp. 287–288, October 2002.

[20] H. Kirchhoff, S. Dixon, and A. Klapuri, "Cross-recording adaptation of musical instrument spectra," Tech. Rep. C4DM-TR-11-12, Queen Mary University of London, 2012, Available from: `http://www.eecs.qmul.ac.uk/~holger/C4DM-TR-11-12`.

[21] MIREX 2007, "Development set for multiple fundamental frequency estimation & tracking," Available from: `http://www.music-ir.org/evaluation/MIREX/data/2007/multiF0/index.htm`.

[22] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, 2010.

[23] Joachim Fritsch, "High quality musical audio source separation," M.S. thesis, UPMC / IRCAM / Telecom ParisTech, 2012, Trios dataset available from: `http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/20`.