2011-6

User Assisted Source Separation Using Non-negative Matrix Factorisation

Derry Fitzgerald
Technological University Dublin, derry.fitzgerald@tudublin.ie

Funder: Science Foundation Ireland

User Assisted Source Separation using Non-negative Matrix Factorisation

Derry FitzGerald[†]

Audio Research Group
School of Electrical Engineering Systems
Dublin Institute of Technology
Kevin St, Dublin 2

E-mail: [†]derry.fitzgerald@dit.ie

Abstract — **Much research has been carried out on the use of non-negative matrix factorisation for the purpose of musical sound source separation. However, a notable shortcoming of non-negative matrix factorisation is that the recovered basis functions have to be clustered to sound sources for separation to take place. This has proved to be a difficult problem to solve. As a means of overcoming this problem, we introduce an extension to non-negative matrix factorisation which allows a user to guide the separation by singing, or playing along with, the source they want to separate. This is done through the use of gamma-chain priors. Examples of user assisted separation are also presented.**

Keywords — **Sound Source Separation, Non-negative Matrix Factorisation**

I Introduction

In recent years, Non-negative matrix factorisation (NMF) [1] and various extensions thereof, has been used as a means of attempting sound source separation [2, 3]. Non-negative matrix factorisation attempts to factorise a non-negative matrix $\mathbf{X}$ of size $n \times m$ into matrix factors $\mathbf{A}$ and $\mathbf{S}$:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{AS} \tag{1}$$

where $\mathbf{A}$ is of size $n \times r$, $\mathbf{S}$ is of size $r \times m$ and r is the rank of the factorisation. NMF has found widespread use due to its ability to give a parts-based representation of data sets. Typically factorisation is performed using the generalised Kullback-Liebler divergence as a cost function:

$$D(\mathbf{X}, \hat{\mathbf{X}}) = \sum \mathbf{X} \log \frac{\mathbf{X}}{\hat{\mathbf{X}}} - \mathbf{X} + \hat{\mathbf{X}} \tag{2}$$

where summation takes place over all elements of $\mathbf{X}$ and $\hat{\mathbf{X}}$. Multiplicative update equations for $\mathbf{A}$ and $\mathbf{S}$ can then be derived from the cost function.

In the case of musical audio signals, $\mathbf{X}$ is typically a magnitude spectrogram, such as obtained via the Short Time Fourier Transform (STFT) and $\mathbf{A}$ contains a set of frequency basis functions which typically correspond to notes or chords played in the signal. $\mathbf{S}$ contains the corresponding time activations of the frequency basis functions which show when the notes or chords are playing. The principal shortcoming of the standard NMF algorithm for sound source separation lies in the fact that the recovered basis functions have to be clustered to their respective sources before the separated signals can be recovered.

The clustering of the basis functions to sources has proved a difficult problem to solve, and it is only in the past two years that progress has been made in solving this problem directly. Spiertz et al [4] make use of log Mel-Frequency Cepstral Coefficients (log MFCCs) obtained from the frequency basis functions. They assume a source-filter model where each note played by a source is the product of a harmonic excitation with an instrument specific filter. They then apply NMF to the log MMCCs in an attempt to learn the filters for a given number of sources. The output of this NMF can then be used to cluster the basis functions.

An alternative approach has been proposed by Jaiswal et al [5]. They make the assumption that, in the log-frequency domain, basis functions that belong to the same source can be shifted up or down in frequency to approximate different notes played by the instrument. The frequency basis functions are transformed from a linear to a log-frequency basis, and then Shifted NMF [6] is performed on these basis functions to learn instrument-specfic basis functions which can be shifted up or down in frequency to approximate the log-frequency basis functions of individual notes played by each instrument. The activations of these instrument basis functions then indicate which of the original basis functions belong to which source.

These algorithms represent considerable advances in the problem of clustering NMF basis functions for sound source separation and have been demonstrated to work on monophonic mixtures containing 2 or 3 pitched instruments. Nevertheless, there is still considerable work to be done before the clustering problem can be considered solved as these approaches do not scale well to more complicated mixtures.

Other NMF-based approaches have tried to overcome the clustering problem by avoiding it altogether through the incorporation of additional constraints into the NMF framework such as shift invariance in frequency, source-filter modeling, harmonicity and temporal continuity [7, 8]. Again these approaches have been demonstrated to work well on simple 2 or 3 instrument mixtures, but have difficulty with more complex mixtures.

II Gamma Priors

Part of the problem with the above approaches is that they are completely blind and operate without knowledge of the sources to be separated. This can be ameliorated if knowledge of the sources can be provided to the separation algorithm. One approach to the incorporation of such knowledge is through the use of Bayesian extensions to NMF [9]. Here, prior information on the frequency basis functions to be separated is incorporated to the NMF framework through the use of the Gamma distribution:

$$\mathcal{G}(y:\alpha,\beta) = y^{\alpha-1}\beta^{-\alpha}e^{-y/\beta}/\Gamma(\alpha) \tag{3}$$

where the Gamma distribution is defined for $y > 0$. It was then assumed that each entry in the frequency basis functions was drawn independently from a Gamma distribution yielding:

$$\begin{aligned} p(\mathbf{A}_{i,k}) &= \mathcal{G}(\mathbf{A}_{i,k}:\alpha_{i,k},\beta_{i,k}^{-1}) \\ &= \mathbf{A}_{i,k}^{\alpha_{i,k}-1}\beta_{i,k}^{\alpha_{i,k}}e^{\mathbf{A}_{i,k}\beta_{i,k}}/\Gamma(\alpha_{i,k}) \end{aligned} \tag{4}$$

The hyperparameters $\alpha_{i,k}$ and $\beta_{i,k}^{-1}$ can be chosen independently for each frequency basis function in $\mathbf{A}_{1:n,k}$. A simple interpretation of $\beta_{i,k}^{-1}$ is as a set of weights which describe the typical or expected frequency spectrum of a given source, such as the typical spectrum of a snare drum for example. Alternatively, in the case of pitched instruments $\beta_{i,k}^{-1}$ could be chosen so that the gamma priors have peaks at harmonically related frequencies. In effect, the gamma priors are used to push the frequency basis functions to have a set of desirable characteristics related to the sources to be separated.

The standard NMF cost function can then be extended to incorporate the gamma priors, yielding an extended cost function:

$$D(\mathbf{X},\hat{\mathbf{X}}) + \log(p(\mathbf{A}_{i,k})) \tag{5}$$

from which update equations can be derived. Gamma priors have successfully been used to perform separation of pitched sources from percussion sources [10]. The use of gamma priors as described above can also be easily extended to incorporate priors on the time activations, though this was not done by Virtanen et al, who instead imposed a temporal continuity constraint on the time activation functions.

III User Assisted Source Separation

Recently, Smaragdis has proposed a user assisted separation technique for monophonic source separation in the context of Probabilistic Latent Component Analysis [11]. Here the user hummed or sang along in time with the monophonic mixture, with the user vocally approximating the source they wished to separate or isolate from the mixture. The user guide signal was recorded and then decomposed and used as a guide to separate out the source of interest. Here we propose to perform a similar user assisted analysis, but in the context of the widely used NMF framework.

The user guide signal is recorded and transferred to the time-frequency domain using an STFT from which a magnitude spectrogram $\mathbf{X}_u$ is obtained. This magnitude spectrogram is then decomposed using a standard NMF algorithm, yielding a set of frequency and time basis functions $\mathbf{A}_u$ and $\mathbf{S}_u$ respectively.

Similarly, a magnitude spectrogram $\mathbf{X}$ is obtained from the actual mixture signal. This is then decomposed using NMF with gamma priors. It is assumed that a subset of the basis functions to be obtained from NMF correspond to the source to be separated and so $\mathbf{A}_u$ is used as as a set of hyperparameters for the gamma priors on this subset of the frequency basis functions. Similarly $\mathbf{S}_u$ is used as a set of hyperparameters on a gamma distribution over the corresponding subset of time basis functions. The remaining frequency and time basis

functions have no priors applied to them and are free to adapt to capture the remaining sources.

This can potentially overcome the problem of clustering the basis functions for the source of interest, as the system will know which subset of basis functions has been adapted to match the source. This subset of basis functions can then be used to reconstruct the source without recourse to clustering. This is the principal advantage of user assisted source separation, in that the information provided by the user guide signal is used to guide NMF towards a factorisation in which a subset of the basis functions have a set of desired frequency and temporal characteristics.

However, there will be a mismatch between the characteristics of the source to be separated, and the characteristics of the user guide signal. Therefore, it is proposed that the gamma priors be used initially to point the basis functions towards the source of interest. Then, as the number of iterations increases, the effects of the gamma priors are gradually reduced to zero, thereby allowing the subset of the basis functions adapt to the characteristics of the source as found in the mixture signal as opposed to that of the guide signal. Once the gamma priors have been reduced to zero, the update equations then collapse to those of standard NMF, and it is expected that by this time the guided basis functions will have been guided to the desired set of characteristics.

In order to derive update equations which take into account both frequency and temporal priors, it is necessary to extend the cost function in eqn 5 as follows:

$$C = D(\mathbf{X}, \hat{\mathbf{X}}) + \log(p(\mathbf{A}_{i,k})) + \log(p(\mathbf{S}_{k,l})) \tag{6}$$

where C denotes the extended cost function and where l indexes from over the time frames in the magnitude spectrogram from $1 : m$.

Multiplicative update equations can then be derived for the variables of the model. These update equations take the form:

$$\mathbf{R} = \mathbf{R} \otimes \frac{\bigtriangledown^{-}_{\mathbf{R},C}}{\bigtriangledown^{+}_{\mathbf{R},C}} \tag{7}$$

where $\mathbf{R}$ denotes the variable in the model to be updated and where $\bigtriangledown^{-}_{\mathbf{R},C}$ and $\bigtriangledown^{+}_{\mathbf{R},C}$ denotes the negative and positive parts respectively of the partial derivative of C with respect to $\mathbf{R}$. $\otimes$ denotes elementwise multiplication and all divisions in the above and subsequent equations are taken as elementwise.

The resulting update equations for $\mathbf{A}$ and $\mathbf{S}$ are then as follows:

$$\mathbf{A} = \mathbf{A} \otimes \frac{\lambda_{\mathbf{A}}((\alpha_{\mathbf{A}} - 1)\mathbf{A}) + (\mathbf{X}/\hat{\mathbf{X}})\mathbf{S}'}{\lambda_{\mathbf{A}}\beta_{\mathbf{A}} + \mathbf{O}\mathbf{S}'} \tag{8}$$

$$\mathbf{S} = \mathbf{S} \otimes \frac{\lambda_{\mathbf{S}}((\alpha_{\mathbf{S}} - 1)\mathbf{S}) + \mathbf{A}'(\mathbf{X}/\hat{\mathbf{X}})}{\lambda_{\mathbf{S}}\beta_{\mathbf{S}} + \mathbf{A}'\mathbf{O}} \tag{9}$$

where $\mathbf{O}$ is an all-ones matrix of size $n \times m$ and $'$ denotes matrix transpose. In line with the settings proposed by Virtanen in [9],all elements in $\alpha_{\mathbf{A}}$ and $\alpha_{\mathbf{S}}$ are set to a value of 1. $\lambda_{\mathbf{A}}$ and $\lambda_{\mathbf{S}}$ are parameters which control the influence of the gamma priors on the factorisation for $\mathbf{A}$ and $\mathbf{S}$ respectively. These are initially set at 1 and gradually decreased to 0 over a number of iterations. Finally, $\beta_{\mathbf{A}}$ and $\beta_{\mathbf{S}}$ are set as follows:

$$\beta_{\mathbf{A}} = 1/\mathbf{A}_u \tag{10}$$

$$\beta_{\mathbf{S}} = 1/\mathbf{S}_u \tag{11}$$

where as previously noted, $\mathbf{A}_u$ and $\mathbf{S}_u$ are the frequency and time basis functions obtained by performing NMF on a magnitude spectrogram of the user guide signal.

The update equations for the unguided basis functions can easily be obtained by setting $\lambda_{\mathbf{A}}$ and $\lambda_{\mathbf{S}}$ to zero, resulting in the standard update equations for NMF, and it can be seen that the guided factorisation also collapses to standard NMF updates once the effects of the gamma priors has been removed.

On completion of the factorisation, the target source can be recovered from:

$$\mathbf{X}_t = \mathbf{A}_t\mathbf{S}_t \tag{12}$$

where $\mathbf{X}_t$ is the recovered magnitude spectrogram of the target source, $\mathbf{A}_t$ contains the frequency basis functions associated with the target source, and $\mathbf{S}_t$ contains the time basis functions associated with the target source. Similarly the basis functions associated with the remaining sources can be recovered from

$$\mathbf{X}_r = \mathbf{A}_r\mathbf{S}_r \tag{13}$$

where $\mathbf{X}_r$ is the recovered magnitude spectrogram of the remaining sources, $\mathbf{A}_r$ contains the frequency basis functions associated with these sources, and $\mathbf{S}_r$ contains the time basis functions associated with the remaining sources.

The original phase information of the STFT of the mixture signal can be applied to these spectrograms to allow resynthesis, however better results can be obtained by using these spectrograms to generate a Wiener filter to apply to the original complex-valued spectrogram of the mixture signal $\mathbf{Y}$:

$$\mathbf{Y}_t = \mathbf{Y} \otimes \frac{\mathbf{X}_t^2}{\mathbf{X}_t^2 + \mathbf{X}_r^2} \tag{14}$$

$$\mathbf{Y}_r = \mathbf{Y} \otimes \frac{\mathbf{X}_r^2}{\mathbf{X}_t^2 + \mathbf{X}_r^2} \tag{15}$$

where all operations are performed elementwise and where $\mathbf{Y}_t$ and $\mathbf{Y}_r$ are complex-valued spectrograms containing the target source and the remaining sources respectively. These can be transformed to time-domain signals using the inverse STFT.

IV User Assisted Separation Examples

The utility of user assisted source separation is now illustrated through a number of real-world examples. The first of these is an excerpt from "God Only Knows" by the Beach Boys. The user was recorded singing along with the lead vocal in this excerpt, and so the algorithm will attempt to recover the lead vocal from the original signal in this case. Both the original signal and guide signal had a sampling rate of 44.1 kHz. Spectrograms were obtained via an STFT with an FFT size of 4096, windowed with a Hamming window of 4096 points, and with a hopsize of 1024. Figure 1 shows the spectrogram of the original mixture signal, and the vocal part can be seen as a set of modulating sinusoids in the spectrogram.

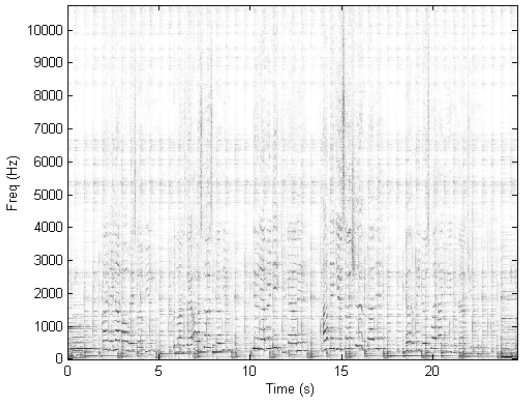

Fig. 1: Spectrogram of excerpt from "God Only Knows".

Figure 2 then shows the spectrogram of the guide vocal. NMF was performed on this spectrogram with NMF set to recover 50 frequency and time basis function pairs, with 100 iterations of the algorithm performed. These basis functions were then used to create gamma priors to guide the separation of the original mixture signal. Here factorisation was performed with 100 frequency and time basis function pairs, 50 of which were guided with the gamma priors, while the remainder were free to adapt to the mixture signal without interference from the priors. $\lambda_{\mathbf{A}}$ and $\lambda_{\mathbf{S}}$ were both initially set to 1 and reduced to 0 over the first 6 iterations, with the total number of iterations again set to 100.

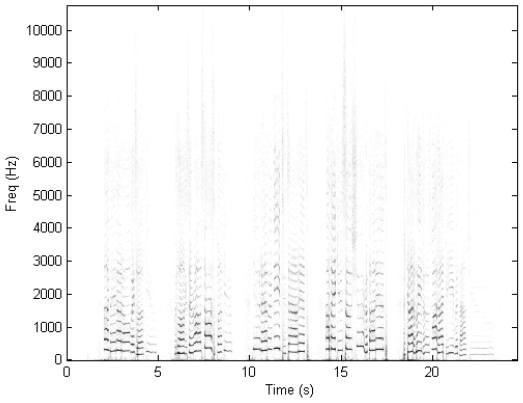

Fig. 2: Spectrogram of guide vocal for "God Only Knows"

Complex valued spectrograms were then obtained as described in the previous section, and the time domain signals recovered via inverse STFT. Figure 3 then shows the spectrogram of the separated vocal obtained from the original mixture. It can be seen that the vocal has successfully been separated from the original mixture signal. Further, the modulations in the vocal match those present in the original mixture spectrogram, in comparison to the extra modulations visible in the guide vocal. This demonstrates that the algorithm has adapted to the characteristics of the vocal in the original signal as opposed to those of the guide vocal. However, it can also be seen that there are still some traces of the other sources, in particular some percussion, evident in the separated spectrogram.

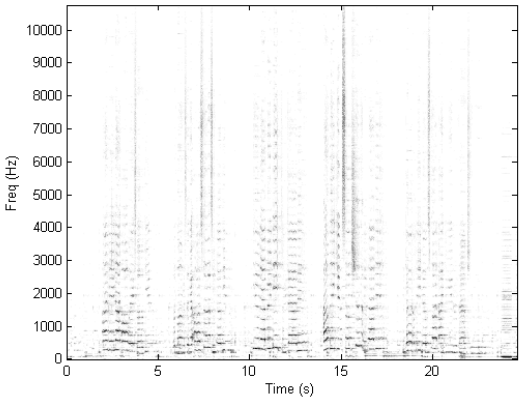

Fig. 3: Spectrogram of vocal separation from "God Only Knows"

Figure 4 shows the spectrogram of the separated instrumental track, and it can be seen that very little trace remains of the vocal in this separation. On listening to the separations, the vocal was observed to be separated very well, with the level of the instrumental track greatly reduced. Similarly,

the presence of the vocal is greatly reduced in the separated instrumental track.

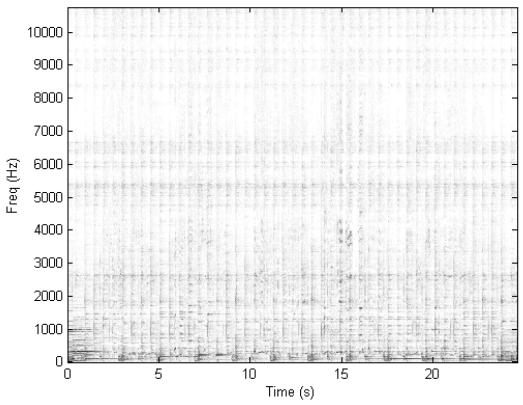

Fig. 4: Spectrogram of instrumental track separated from "God Only Knows".

The previous example dealt with a case where both the frequency and time characteristics of the guide signal were a reasonable match to those of the source to be separated. However, this will not always be the case. For example, the source of interest could be a set of chords played on piano or guitar. The second example is an excerpt from "Photograph" by Def Leppard, containing a lead vocal, guitar and drums. Here the user attempts to separate the guitar chords by singing "doo-dooh" sounds in time with the guitar part. The same settings were used for the separation as those of the previous example.

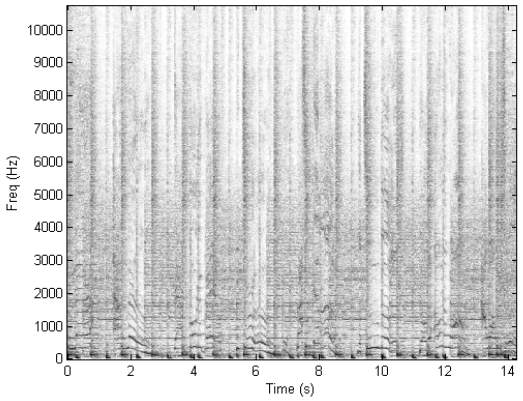

Fig. 5: Spectrogram of excerpt from "Photograph".

Figure 5 shows the spectrogram of the original excerpt from "Photograph". The guitar part can be seen as sets of horizontal lines in the spectrogram, with the drums visible as vertical lines, and the vocal as a set of modulated sinusoids. Figure 6 shows the spectrogram of the user singing along with the guitar part, while figure 7 shows the spectrogram of the recovered guitar part. Finally, figure 8 shows the separated vocal and drum parts.

It can be seen that the spectral characteristics of the guide source are considerably different from those of the separated guitar part, while the temporal information is broadly similar. Despite the difference in spectral characteristics, enough information has been provided to the algorithm to point the guided basis functions towards the guitar part. However,some traces of the vocal and drums can be seen in the spectrogram, and these traces are also evident on listening to the separated guitar part. Nonetheless, the guitar part clearly predominates in the separated signal. In the separated vocal and drums track, the level of the guitar part has been greatly reduced, though some traces of it can still be heard.

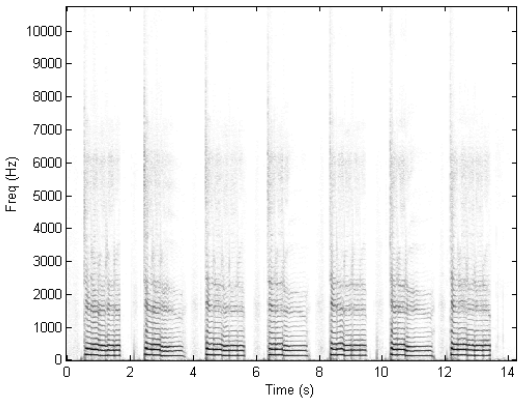

Fig. 6: Spectrogram of guide vocal for the guitar part of "Photograph".

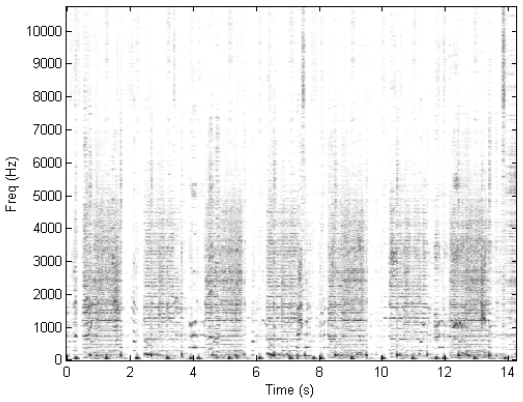

Fig. 7: Spectrogram of separated guitar part from "Photograph".

The above examples demonstrate that user assisted source separation using NMF is a viable technique for separating sources from monophonic mixtures of sounds, and that the technique can handle different types of sources, even when there

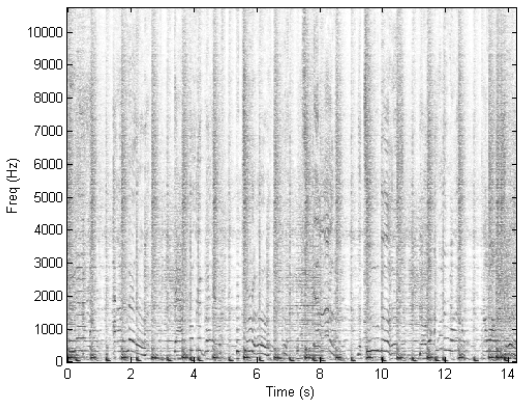

Fig. 8: Spectrogram of vocals and drums from "Photograph".

is considerable mismatch between the spectral characteristics of the guide source and the source to be separated. The above examples can be heard at `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=53`

V Conclusions

Having outlined the use of NMF for the purposes of sound source separation, the problem of clustering the recovered basis functions to sources was highlighted. A number of directions in attempting to solve this problem were discussed, before the concept of user assisted source separation was proposed as a means of overcoming the clustering problem associated with NMF. An algorithm to perform user assisted sound source separation was then proposed, with the user provided information incorporated into the NMF framework by means of gamma priors. The utility of this approach was then demonstrated through a number of real-world examples. Future work will concentrate on extending this technique to handle multichannel signals.

References

[1] D. Lee, and H. Seung, Algorithms for non-negative matrix factorization, Adv. Neural Info. Proc. Syst. 13, 556-562 (2001).

[2] T. Virtanen, "Sound Source Separation in Monaural Music Signals", Tampere University of Technology, 2006.

[3] P. Smaragdis, "Non-Negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs". In 5th International Conference on Independent Component Analysis and Blind Signal Separation, Grenada, Spain, September 2004.

[4] M. Spiertz and V. Gnann, Source-filter based clustering for monaural blind source separation, in Proc. of International Conference on Digital Audio Effects DAFx '09, (Como, Italy), Sept. 2009.

[5] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle and S. Rickard, "Clustering NMF basis functions using shifted NMF for Monaural Sound Source Separation" IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, 2011.

[6] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted Non-negative Matrix Factorisation for Sound Source Separation", Proceedings of the IEEE conference on Statistics in Signal Processing, Bordeaux, France, July 2005.

[7] D. FitzGerald, M. Cranitch,and E. Coyle, Extended Nonnegative Tensor Factorisation Models for Musical Sound Source Separation, Computational Intelligence and Neuroscience, 2008.

[8] A. Ozerov, E. Vincent, F. Bimbot, "A general modular framework for audio source separation", 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10), Sep. 27, 2010, Saint-Malo, France

[9] T. Virtanen, A. T. Cemgil, and S. J. Godsill. Bayesian Extensions to Non-negative Matrix Factorisation for Audio Signal Modelling, IEEE International Conference on Acoustics, Speech and Signal Processing, 2008

[10] D. FitzGerald, E. Coyle, and M. Cranitch, "Using Tensor Factorisation Models to Separate Drums from Polyphonic Music", Proceedings of the International Conference on Digital Audio Effects (DAFX09), Como, Italy, 2009

[11] P. Smaragdis and G. Mysore. 2009. Separation by humming: User-guided sound extraction from monophonic mixtures. In proceedings of IEEE Workshop on Applications Signal Processing to Audio and Acoustics. New Paltz, NY. October 2009.