2012

On Inpainting the Adress Algorithm

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

Dan Barry
Technological University Dublin

Recommended Citation

Fitzgerald, D. & Barry, D. (2012) On inpainting the adress olgorithm. *23rd IET Irish Signals and Systems Conference*, Maynooth, Ireland. June 28-29th. 2012

Funder: SFI

On Inpainting the Adress Algorithm

Derry FitzGerald[†] and Dan Barry[*]

Audio Research Group
Dublin Institute of Technology,
Kevin St, Dublin 8, IRELAND

E-mail: [†]derry.fitzgerald@dit.ie [*]dan.barry@dit.ie

Abstract — **The Adress algorithm has been demonstrated to be capable of separating sound sources from instantaneous linear mixtures, provided that the sources have a unique pan position in the stereo field. However, a shortcoming of the Adress algorithm is that all time-frequency bins outside of the chosen azimuth range are set to zero, resulting in audible artifacts in the resynthesised sound. Here we show that an inpainting algorithm based on NMF is capable of estimating these missing values and improves on the results obtained using Adress only.**

Keywords — **Sound Source Separation, Audio Inpainting**

I Introduction

Audio inpainting has been proposed as a means of estimating missing values or corrupted values in audio signals [1]. Typical applications include the recovery of portions of audio distorted by impulsive noise or clipping, or the estimation of missing portions of a streamed audio signal due to to packet loss. In the case of distorted audio it is assumed that the location of the distortions are known and the actual data in these locations is then assumed to be missing.

A number of different approaches have been suggested in tackling the audio inpainting problem. Adler et al propose a time domain approach where the signal is split into overlapping time domain frames, and the missing frames are then estimated using a dictionary-based Orthogonal Matching Pursuit algorithm [1]. Smaragdis et al propose a factorisation-based approach which operates on time-frequency spectrograms of the audio signal [2], as does Le Roux et al, who use a convolutive factorisation model in conjunction with sparsity constraints to estimate the missing values [3].

The above algorithms have been demonstrated to give good results in estimating missing values in audio signals in a wide variety of scenarios, but not as yet to a class of source separation algorithms which are a natural fit with the audio inpainting problem. These are the class of separation algorithms which make use of binary masking to separate sources, such as the DUET algorithm [4], or the Adress algorithm [5]. These algorithms allocate time-frequency bins to individual sources based on the estimation of various parameters, resulting in source spectrograms with many time-frequency bins set to zero where the energy related to the actual source present is missing.

It is proposed to investigate the use of audio inpainting as a post-processing stage on the output of these algorithms to see if the quality of the separations obtained can be improved. In particular, we focus on the Adress algorithm, which is described in Section II. We then briefly review non-negative matrix factorisation (NMF) in Section III. Following on from this, we describe an NMF-based audio inpainting method inspired by that of Le Roux et al [3] in Section IV, as well as explain its application to the outputs obtained from the Adress algorithm. In Section V the use of audio inpainting is illustrated on real-world examples. Finally in section VI conclusions are drawn and areas for future work highlighted.

II The Adress Algorithm

Since the advent of multi-channel recording systems, most popular music recordings are made by

recording the various sources individually. These sources are then electronically summed and distributed across 2 channels using a mixing console. Localisation (or panning) of a source at a given point in the stereo field is achieved by means of a panoramic potentiometer, which divides a given sound source across 2 channels with continuously variable intensity ratios. Increasing the gain in one channel against that of the other channel gives the appearance that the source is localised more in that channel. It should be noted that in this case the phase of the source in both channels is identical and that only the intensity differs. It is this fact which the Adress algorithm utilises in order to perform sound source separation.

The Adress algorithm performs sound source separation on linear instantaneous stereo mixtures of audio signals. It assumes that each source occupies a unique point in the stereo field and separates sources based on their pan position [5]. The linear instantaneous mixing model used in Adress is given by:

$$L(t) = \sum_{j=1}^{J} Pl_j S_j(t) \tag{1}$$

$$R(t) = \sum_{j=1}^{J} Pr_j S_j(t) \tag{2}$$

where S_j indicates the jth source, Pl_j and Pr_j, the panning coefficients for the jth source, and L and R indicate the left and right channel mixtures respectively. Then, an intensity ratio for each source can be defined as:

$$I_j = \frac{Pl_j}{Pr_j} \tag{3}$$

Due to the linear instantaneous mixing model, it can be seen that $L - I_j R$ will cancel the jth source from the mixture. However this will not allow recovery of the cancelled source and so recovery of the cancelled source is done using frequency domain techniques.

To achieve source recovery, a Short Time Fourier Transform (STFT) is carried out on each of the two mixture signals. We define β as the azimuth resolution, which determines how many equally spaced gain scaling values are used to create the frequency-azimuth plane defined across the full stereo space. As the intensity ratio I is not bounded, we define a bounded gain scale vector g. For a given azimuth resolution, the gains g are defined as:

$$g_i = \begin{cases} \dfrac{i}{\beta} & \text{if} \quad i \leq \beta/2 \\ \dfrac{\beta - i}{\beta} & \text{if} \quad i > \beta/2 \end{cases} \tag{4}$$

where $0 \leq i \leq \beta$ and where i and β are integers. Similarly we define a position index

$$P_i = \begin{cases} g_i - 1 & \text{if} \quad i \leq \beta/2 \\ 1 - g_i & \text{if} \quad i > \beta/2 \end{cases} \tag{5}$$

The values of P then range from -1 for sources panned hard left, to 1 for sources panned hard right, with 0 indicating a position in the centre. The values of g then range from 0 for sources hard left, increasing to 1 for centre positioned sources, before decreasing to 0 for sources panned hard right.

The frequency-azimuth plane is then defined as:

$$Az_{k,i} = \begin{cases} |Lf_k - g_i Rf_k| & \text{if} \quad i \leq \beta/2 \\ |Rf_k - g_i Lf_k| & \text{if} \quad i > \beta/2 \end{cases} \tag{6}$$

where Rf_k and Lf_k denote the kth frequency bin of the current right and left frames of the STFT respectively.

In order to resynthesise a given source, it is necessary to define a source position d, which is a value taken from P. When a source occurs at this source position, the energy in the frequency bins associated with a given source will be cancelled out. This results in a minimum at that position in the azimuth frequency plane. This minimum then contains the residual energy present due to other sources in the mixture. However, due to frequency overlap between different sources, the position of a given frequency minimum can move away from that of the actual source position. In order to overcome this problem, an azimuth subspace width, H is defined, so that $-1 \leq H \leq 1$. Together with d, this defines which azimuth positions in P are to be used for resynthesis. The source spectrogram for the current frame can then be estimated from

$$Y_k = \begin{cases} E & \text{if} \quad d - H/2 \leq \text{argmin}(Az_{ki}) \leq d + H/2 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where E is defined as:

$$E = \begin{cases} Lf_k - \min(Az_{ki}) & \text{if} \quad d \leq 0 \\ Rf_k - \min(Az_{ki}) & \text{if} \quad d > 0 \end{cases} \tag{8}$$

The phase information from the channel in which the source is dominant can then be applied to this spectrogram to allow resynthesis in the time domain via an inverse STFT once all the frames have been estimated.

It can be seen from the above that there is a trade-off involved in choosing the azimuth subspace width. Widening the width allows the capture of greater numbers of bins related to the source of interest, but at the expense of increasing

the likelihood of capturing bins belonging to other sources present in the mixture, thereby increasing the amount of bleed present and the amount of artefacts present. In any case, for bins in which another source predominates in terms of energy, it is likely that these bins will fall outside the chosen azimuth subspace, resulting in an energy of zero for that bin even if some energy due to the source of interest is present. Therefore, it can be seen that a means of estimating the source energy at those bins would be advantageous in improving the quality of the separated source. Audio inpainting is one such means of estimating this missing information.

III NMF

Non-negative matrix factorisation (NMF) [6] and extensions of NMF have been widely used to attempt sound source separation [7, 8]. In the case of audio signals NMF attempts to factorise a non-negative spectrogram, such as a magnitude spectrogram obtained via an STFT, $\mathbf{X}$ of size $n \times m$ into matrix factors $\mathbf{A}$ and $\mathbf{S}$:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{AS} \tag{9}$$

where $\mathbf{A}$ is of size $n \times r$ containing a set of frequency basis functions, $\mathbf{S}$ is of size $r \times m$ containing a set of corresponding time basis functions, and where r is the rank of the factorisation. This yields a parts-based decomposition where the frequency basis functions typically contain frequency spectra corresponding to events such as notes, chords or drum hits played during the piece of music. The time basis functions then provide information relating to when these events occurred.

A commonly used cost function for performing NMF is the generalised Kullback-Liebler divergence:

$$D(\mathbf{X}, \hat{\mathbf{X}}) = \sum \mathbf{X} \log \frac{\mathbf{X}}{\hat{\mathbf{X}}} - \mathbf{X} + \hat{\mathbf{X}} \tag{10}$$

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 yielding:

$$\mathbf{A} = \mathbf{A} \otimes \frac{(\mathbf{X}/\hat{\mathbf{X}})\mathbf{S}^\intercal}{\mathbf{O}\mathbf{S}^\intercal} \tag{11}$$

$$\mathbf{S} = \mathbf{S} \otimes \frac{\mathbf{A}^\intercal(\mathbf{X}/\hat{\mathbf{X}})}{\mathbf{A}^\intercal\mathbf{O}} \tag{12}$$

where $\mathbf{O}$ is an all-ones matrix of size $n \times m$ and $^\intercal$ denotes matrix transpose. All divisions are taken as elementwise and $\otimes$ denotes elementwise multiplications.

IV NMF-based Audio Inpainting

In this paper, we are not interested in using NMF for the purpose of sound source separation, but instead to estimate missing values in an audio spectrogram, in this case obtained as the output of the Adress algorithm. We intend to take advantage of properties of the Adress algorithm in conjunction with the linear parts-based decomposition obtained from NMF to estimate the missing values in the Adress spectrogram. The position of a given bin in the frequency-azimuth plane is a function of the sources that present in that bin, and so the position of a given frequency will change with time as sources come in and out. Taking a guitar chord as an example, each time the chord is played, it will typically be overlapped at different frequencies as the vocal melody changes in pitch or different drum sounds are played against the chord. Therefore, different parts of the chord spectra will fall within the azimuth subspace each time the chord is played. In other words, the separation obtained via Adress varies locally with time.

In contrast, NMF gives a global linear decomposition of the entire spectrogram, capturing repeating parts across the whole signal. This means that when the repeating guitar chord occurs, even if different parts of the chord spectrum are present at different times, then NMF will attempt to find a basis function which captures the global characteristics of the chord, and so parts that are missing at one instance of the chord will tend to be filled in by parts that are present at another instance, leading to a basis function which is closer to the actual frequency characteristics of the chord than any of the individual occurrences in the Adress spectrogram.

Therefore, simply performing NMF on the spectrogram $\mathbf{X}$ output by the Adress algorithm , should result in an improved estimate of the source, estimated from $\hat{\mathbf{X}} = \mathbf{AS}$. However, this can sometimes introduce noise, especially in bins which had low energy to begin with. To help eliminate this noise, we replace any bins in the estimated spectrogram which have energy greater than that of the mixture spectrogram with the bins from the mixture spectrogram in which the source is dominant. Assuming that the bins estimated originally using Adress are accurate, this spectrogram can be further improved by substituting the original non-zeros values of $\mathbf{X}$ back into $\hat{\mathbf{X}}$.

However, this is still not an ideal way of estimating the missing values. This is because the NMF basis functions will still try to take into account the zeros in $\hat{\mathbf{X}}$ and so try to suppress bins where data should be present. Therefore, we adapt the approach taken in [3], where they estimated a convolutive NMF model from a spectrogram containing missing columns. After an initial factorisation on the original spectrogram with the missing data, the output from the convolutive NMF model was used to fill in the values missing from the original spectrogram. A second convolutive NMF factorisation was then performed, but at each iteration, the missing values were then updated using the

values estimated from the previous iteration of the factorisation. This has the effect of allowing the missing values to converge over the iterations to values which are consistent with the factorisation of the observed data, and so these values are now more likely to be good estimates of the actual missing data.

Also included by Le Roux et al was a sparsity prior on the time basis functions. This was included in an attempt to prevent the inpainting algorithm from putting energy into bins where no energy was actually present. The sparsity prior pushes the time basis functions to be as sparse as possible while still giving a good reconstruction of the original data, and so acts as a brake to prevent the inpainting algorithm from incorporating energy where none should be present.

When performing inpainting on the outputs from Adress, we do not use the convolutive NMF algorithm used by Le Roux et al. The convolutive NMF algorithm used is particularly suited to modeling solo pitched instruments, but is not effective at capturing transients associated with percussion sounds, or indeed the transients associated with the onset of certain pitched instruments such as piano or guitar. Further, the convolutive NMF model was tested by eliminating groups of columns from the audio spectrogram, and so needed the harder constraints imposed by convolutive NMF to recover the missing frames. This is as opposed to the Adress output which contains no missing columns, and as already noted, due to the nature of the Adress algorithm, some information related to the source will always be present, allowing a standard NMF decomposition to recover the sources.

We incorporate a sparsity prior on $\mathbf{S}$ through the use of the $L1$ norm. The NMF cost function then becomes:

$$D(\mathbf{X}, \hat{\mathbf{X}}) + \lambda|\mathbf{S}|^1 \tag{13}$$

where λ controls the degree of sparseness of the solution. The update equation for $\mathbf{S}$ is then given by:

$$\mathbf{S} = \mathbf{S} \otimes \frac{\mathbf{A}^\intercal(\mathbf{X}/\hat{\mathbf{X}})}{\mathbf{A}^\intercal\mathbf{O} + \lambda} \tag{14}$$

while the update equation for $\mathbf{A}$ remains as previously defined.

The NMF inpainting algorithm for use with Adress can then be defined as follows below. Let J be the non-zero indices in $\mathbf{X}$:

$$J = \arg(\mathbf{X} > 0) \tag{15}$$

Then perform standard NMF on $\mathbf{X}$ to yield initial estimates of $\hat{\mathbf{X}}$, $\mathbf{A}$ and $\mathbf{S}$. Letting p indicate the iteration number, where $p = 0$ denotes the initial estimate obtained via standard NMF, then iterate through the following:

Step 1: $\mathbf{X}^{p+1}_{n,m} = \begin{cases} \hat{\mathbf{X}}^p_{n,m} & \text{if } (n,m) \notin J \\ \mathbf{X}_{n,m} & \text{if } (n,m) \in J \end{cases}$

Step 2: Update $\mathbf{A}$ and $\mathbf{S}$ using eqns. 11 and 14

This process continues until convergence or for a fixed number of iterations. The final value of $\mathbf{X}$ is then used as the inpainted estimate of the source. Despite the presence of the sparsity prior, there will still be additional noise present in the inpainted spectrogram. In order to suppress this to some extent, we search for bins in $\mathbf{X}$ which have energy greater than that in $\mathbf{Y}$ where $\mathbf{Y}$ is the original mixture spectrogram of the channel in which the source being separated is dominant. These values from $\mathbf{Y}$ are then used to replace those in $\mathbf{X}$ as follows:

$$\mathbf{X}_{n,m} = \begin{cases} \mathbf{Y}_{n,m} & \text{if } \mathbf{X}_{n,m} > \mathbf{Y}_{n,m} \\ \mathbf{X}_{n,m} & \text{otherwise} \end{cases} \tag{16}$$

It should be noted that while the use of inpainting can improve the separations obtained from Adress, the use of inpainting still carries with it limitations from Adress. Firstly, sources panned to the same position will still be separated together, and secondly, there is still a trade-off related to the choice of azimuth width between source recovery and bleed from other instruments. However, the use of inpainting ameliorates this trade-off to some extent, allowing a narrower width to be used, reducing bleed from the other sources while still allowing good recovery of the source. However, too small a width results in the introduction of extraneous noise to the recovered signal, as there is no longer enough information for the inpainting algorithm to function properly.

V Inpainting Examples

In this section we demonstrate how the use of inpainting improves the results obtained by the Adress algorithm. Figure 1 shows the spectrogram of a stereo mixture of a song containing guitar, drums, bass, synthesiser, piano and vocals. The guitar was mixed to be dominant in the left channel and so the spectrogram of this channel is shown. The stereo mixture was obtained by mixing the mono sources with the sources spread evenly across the stereo space. This means that there was a distance of 0.4 in the azimuth plane between the sources. Both standard NMF and the inpainting NMF algorithm were ran for 500 iterations, while λ was set to 1000 to encourage sparsity in the output of inpainting NMF.

Figure 2 then shows the spectrogram for the original unmixed guitar signal, while figure 3 then

shows the spectrogram recovered using Adress, where the correct source position was provided to the Adress algorithm. An azimuth width of 0.4 was chosen as this ensured that the azimuth range covered half the distance to the next source. It can be seen that the general time-frequency characteristics of the source have been recovered, though many harmonics are missing, and gaps in many harmonics are evident. On listening to the recovered source, it can be clearly identified as a guitar. Figure 4 shows the guitar spectrogram recovered by using NMF-based inpainting on the output from Adress. Here, it can be seen that more of the harmonics have successfully been recovered, though some of the low frequency harmonics are still missing. There is a noticeable improvement in resynthesis quality when compared to that of Adress.

Figure 5 then shows the original drum spectrogram from the same excerpt, with figure 6 showing the drum spectrogram recovered via Adress, using the correct source position and the same azimuth width as the previous example. Finally, figure 7 shows the drum spectrogram recovered using inpainting. It can be seen that more of the energy of the drums has been recovered after inpainting. On listening to the separated signals, the transients on the signal recovered via inpainting are noticeably sharper than those obtained via Adress, and there is more energy present in the drum sounds.

The inpainting algorithm was also tested on real-world commercial recordings, where the original sources were unavailable for comparison. Here, informal listening tests suggest that the results obtained via the inpainting algorithm again improved upon those obtained via Adress, highlighting the usefulness of the inpainting algorithm for use in real-world separation tasks. Examples of real-world separations can be found at [9].

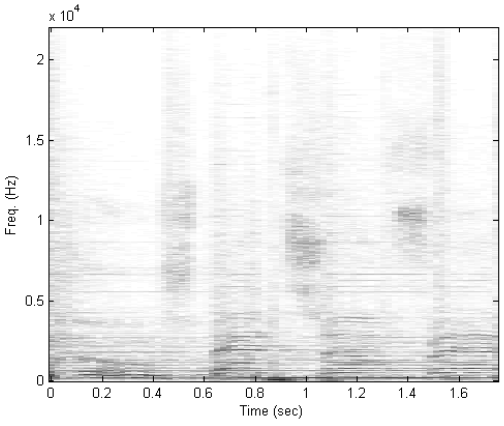

Fig. 1: Spectrogram of left channel of mixture signal

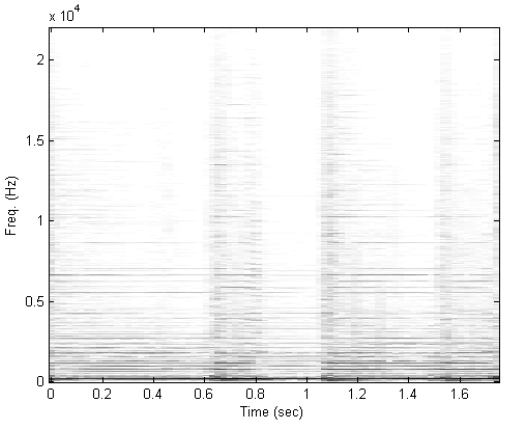

Fig. 2: Spectrogram of original guitar signal

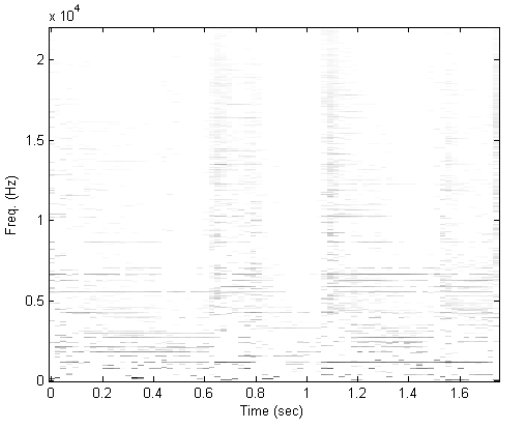

Fig. 3: Spectrogram of guitar recovered using Adress

VI Conclusions

Having given an overview of the Adress algorithm, we then highlighted issues associated with the algorithm. In particular, the nature of Adress means that many bins which should have energy related to the source to be separated will have zero energy. An NMF-based inpainting algorithm was then proposed as a means of estimating these missing values. The improvement in the quality of the separations obtained when inpainting was used with Adress was then demonstrated through real world examples. Future work will concentrate on extending the NMF model to incorporate additional constraints such as temporal continuity with a view to further improving the separations obtained and also on investigating the use of inpainting with other algorithms such as DUET.

VII Acknowledgements

Derry FitzGerald was supported in this research by the Science Foundation Ireland Stokes Lectureship programme.

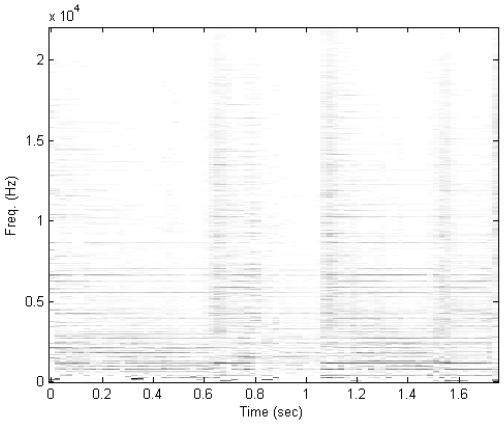

Fig. 4: Spectrogram of guitar estimated using NMF inpainting after Adress

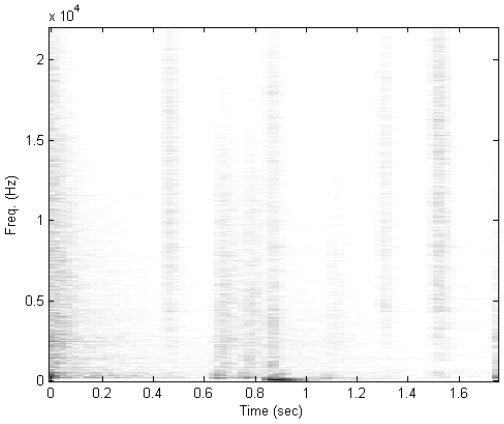

Fig. 5: Spectrogram of original drums signal

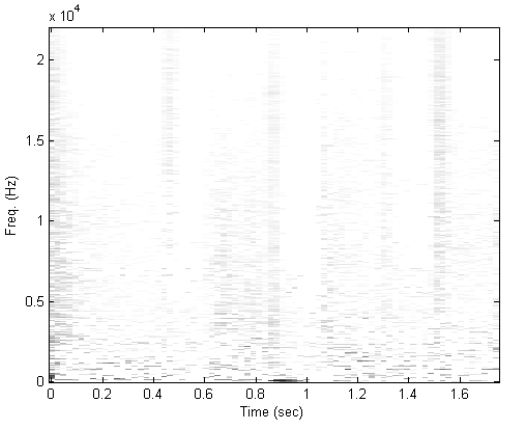

Fig. 6: Spectrogram of drums recovered using Adress

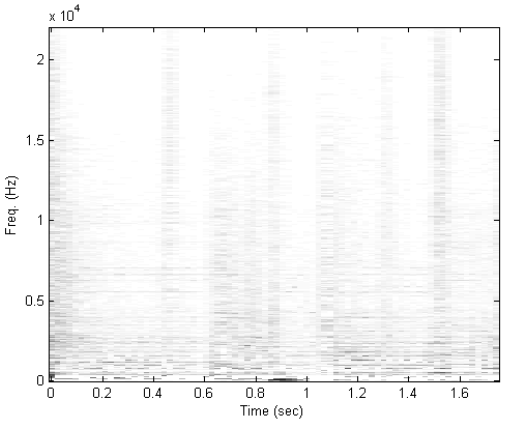

Fig. 7: Spectrogram of drums estimated using NMF inpainting after Adress

References

[1] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley. "Audio Inpainting". *IEEE Trans. on Audio Speech and Signal Processing*, Vol. 20, No. 3, 2012 922 - 932.

[2] P. Smaragdis, B. Raj, and M. Shashanka. "Missing Data Imputation for Time-Frequency Representations of Audio Signals". *Journal of Signal Processing Systems*, 2010 361-370.

[3] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigne, S. Sagayama. "Computational auditory induction as a missing-data model-fitting problem with Bregman divergence". *Speech Communication*, 53 (2011) 658-676.

[4] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking", *IEEE Transactions on Audio Speech and Signal Processing*, Vol. 52, No. 7, 1830-1847 2004.

[5] D. Barry, E. Coyle, and B. Lawlor, "Sound Source Separation: Azimuth Discrimination and Resynthesis", *Proc. 7th International Conference on Digital Audio Effects*, 240-244, 2004

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

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

[8] P. Smaragdis, "Non-Negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs", *5th International Conference on Independent Component Analysis and Blind Signal Separation*,494-499, 2004.

[9] Inpainting Examples `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=62`.