2020-06-01

Remedying Sound Source Separation via Azimuth Discrimination and Re-synthesis

Ruairí de Fréin
Dublin Institute of Technology, ruairi.defrein@dit.ie

Recommended Citation

R. de Fréin, "Remedying Sound Source Separation via Azimuth Discrimination and Re-synthesis," *2020 31st Irish Signals and Systems Conference (ISSC)*, Letterkenny, Ireland, 2020, pp. 1-6, doi: 10.1109/ISSC49989.2020.9180181.

Remedying Sound Source Separation via Azimuth Discrimination and Re-synthesis

R. de Fréin
Technological University Dublin,
Ollscoil Teicneolaíochta Bhaile Átha Cliath,
Ireland
web: https://robustandscalable.wordpress.com

in: 2020 31st Irish Signals and Systems Conference (ISSC). See also BibTeX entry below.

BibTeX:

```
@article{deFrein20Remedying,
  author={R. {de Fr\'{e}in}},
  booktitle={2020 31st Irish Signals and Systems Conference (ISSC)},
  title={Remedying Sound Source Separation via Azimuth Discrimination and Re-synthesis},
  year={2020},
  volume={},
  number={},
  pages={1-6},
  doi={10.1109/ISSC49989.2020.9180181}
  ISBN={978-1-5386-3840-8},
  doi={10.1109/ISSC49989.2020.9180181},
  url={https://ieeexplore.ieee.org/document/9180181},
  keywords={pan-mixing process, time-frequency component,
  three-source music mixtures, source estimates, sound source separation,
  azimuth discrimination, input sound sources, stereo field,
  pan-position, Adress algorithm, stereo mixture,
  inter-aural intensity scaling parameter, azimuth re-synthesis,
  music sound sources, Redress method, audible artefacts},
  }
```


document created on: January 11, 2021
created from file: rdefreinISSC20coversheet.tex
cover page automatically created with CoverPage.sty
(available at your favourite CTAN mirror)

Remedying Sound Source Separation via Azimuth Discrimination and Re-synthesis

Ruairí de Fréin
SEEE, Technological University Dublin, Ireland
rdefrein@gmail.com

***Abstract*—Commercially recorded music since the 1950s has been mixed down from many input sound sources to a two-channel reproduction of these sources. The effect of this approach is to assign sources to locations in a stereo field using a pan-position for each source. The Adress algorithm is a popular way of extracting individual music sound sources from a stereo mixture. A drawback of the Adress algorithm is that when time-frequency components in the stereo mixture are shared between two or more sources, calculating the inter-aural intensity scaling parameter for each source for that time-frequency component is challenging. We show how to obtain a good quality inverse of the pan-mixing process in the time-frequency components which are shared between different sources using a new method called Redress. We demonstrate that we can estimate how much of each source is active in time-frequency components which are shared between sources for two and three-source music mixtures. The consequence of this is that audible artefacts are not as prominent in the source estimates.**

***Index Terms*—Source Separation, Time-Frequency, Music Signal Processing.**

I. Introduction

Since the 1950s commercially recorded music has been mixed down from many input sources to a stereo reproduction of these sources. The effect of this *pan*-mixing approach is to assign sources to locations in a stereo field using a pan-position for each source. The Adress algorithm [1] has gained wide interest due to its success at extracting individual music sound sources from a stereo mixture. It works by using scaling the mixtures on each channel relative to each other to expose nulls across a frequency-gain (frequency-azimuth) plane. Similar to other Time-Frequency (TF) approaches, [2]–[5] source reconstruction is difficult when two or more sources occupy the same TF bin. Our contribution, the Redress algorithm, remedies this short-coming of the Adress algorithm, by de-mixing the contribution of multiple sources to a TF bin by considering frequency-azimuth plane in its entirety. The resulting separated sources exhibit de-mixing artefacts that are less prominent than before.

The Redress algorithm can be classed as a member of the Independent Component Analysis (ICA) family of Source Separation algorithms [6]. ICA techniques are classified according to the problem they solve: the instantaneous mixing problem, the an-echoic mixing problem or the echoic mixing problem. Redress addresses the instantaneous mixing problem. The DUET algorithm [2] and related power weighted relative attenuation and delay estimation approaches [3] are an-echoic de-mixing approaches. Some of these approaches have been extended to the echoic mixing case in [7].

Disjointness, or at least low frequency of overlap, in the TF domain for the constituent sources in a mixture is a desirable property when de-mixing mixtures using the approaches above and [8], [9]. For example, DUET [2] relies on Windowed-Disjoint Orthogonality (WDO), which is a requirement that at most one source is active at any given TF point to successfully separated sources. The WDO assumption is generally sufficiently true for DUET to de-mix mixtures of up to four to five speech sources. Similarly, Non-negative Matrix Factorization (NMF) [9]–[11] works well when sources do not over-lap in a high proportion of the TF bins. The Adress algorithm relies on a similar property for music mixtures. When this property is not evident, the authors described the problem as *smearing* in the frequency-azimuth plane [1]. These ideas have a similar root. Music sources exhibit sparsity in the frequency domain –their energy is concentrated in a few TF bins. When sparsity is coupled with the idea that there is an independence in the occurrence of components, the result is the near-disjointness required by the class of ICA algorithms described above.

What has not be discussed in the literature, is that for many sources mixed using pan-mixing, if the sources are located at different pan-positions, their TF components can be re-expressed as an azimuth trajectory which will have a zero at the position which corresponds to that pan-position. In this paper we show how to represent mixtures using a basis function-trajectory representation; the result of this estimation procedure is that we can separate out the contribution of multiple sources to individual TF bins.

This paper is organized as follows. We introduce a simple pan-de-mixing problem in order to introduce how the Adress algorithm functions in Section II. We introduce the main shorting-coming of the Adress approach using this problem. We then use this de-mixing problem to motivate Redress in Section III. We describe our source reconstruction algorithm in Section IV. Finally, we evaluate the performance of Redress and compare it with Adress in Section V.

II. Mixing model

We start by defining pan-mixing using a specific example as it enables us to establish and define a number of crucial concepts used in the rest of this paper. Two continuous-time sound sources are recorded $s_1(t)$ and $s_2(t)$. They are composed of two frequencies. The first source $s_1(t)$ is composed of frequencies f_1 and f_3 and the second sound source is

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/SIRG/3459.

composed of frequencies f_2 and f_3. We scale both sources by 2 to simplify notation in the remainder of this section,

$$s_1(t) = 2\sin(2\pi f_1 t) + 2\sin(2\pi f_3 t), \tag{1}$$

$$s_2(t) = 2\sin(2\pi f_2 t) + 2\sin(2s\pi f_3 t). \tag{2}$$

Pan-mixing produces a stereo mixture of these sources by weighting the contribution of each source on the left channel and the right channel, the signals $x_1(t)$ and $x_2(t)$. We use two weights $0 < \alpha < 1$ and $0 < \gamma < 1$ and produce

$$x_1(t) = s_1(t) + \alpha s_2(t), \tag{3}$$

$$x_2(t) = \gamma s_1(t) + s_2(t). \tag{4}$$

The Fourier transforms of the original source signals, $s_1(t)$ and $s_2(t)$, are denoted $S_1(f)$ and $S_2(f)$ where f denotes frequency. We may express these transforms compactly

$$\begin{aligned} S_1(f) &= [\delta(f-f_1) + \delta(f+f_1) + \delta(f-f_3) + \delta(f+f_3)] \\ S_2(f) &= [\delta(f-f_2) + \delta(f+f_2) + \delta(f-f_3) + \delta(f+f_3)] \end{aligned} \tag{5}$$

where, $\delta(\cdot)$, is the delta function. The Fourier transforms of the mixtures, $X_1(f)$ and $X_2(f)$, are

$$\begin{aligned} X_1(f) = [&\delta(f-f_1) + \delta(f+f_1) + \alpha\left(\delta(f-f_2) + \right. \\ &\left. \delta(f+f_2)\right) + (1+\alpha)\left(\delta(f-f_3) + \delta(f+f_3)\right)], \end{aligned} \tag{6}$$

$$\begin{aligned} X_2(f) = [&\gamma\left(\delta(f-f_1) + \delta(f+f_1)\right) + \left(\delta(f-f_2) + \right. \\ &\left. \delta(f+f_2)\right) + (1+\gamma)\left(\delta(f-f_3) + \delta(f+f_3)\right)]. \end{aligned} \tag{7}$$

A. Adress

In the Adress algorithm, a frequency-azimuth plane is constructed in order to facilitate the separation of the sources $s_1(t)$ and $s_2(t)$, from the mixtures, $x_1(t)$ and $x_2(t)$. This is achieved by varying an independent variable g over its entire range, $0 \leq g \leq 1$ and computing the magnitude of the difference between the two frequency domain mixtures. To preserve symmetry between the left and the right channels it is necessary to do this scaling for both $X_1(f)$ and $X_2(f)$

$$A_1(f) = |X_1(f) - gX_2(f)|, \tag{8}$$

$$A_2(f) = |X_2(f) - gX_1(f)|. \tag{9}$$

The frequency-azimuth plane that results is denoted $A(f)$ and is produced by concatenating the components, $A_1(f)$ and $A_2(f)$,

$$A(f) = [A_1(f) A_2(f)]. \tag{10}$$

Given this simple mixing problem, we only need to consider three frequency components f_1, f_2 and f_3 and these components can be defined in closed form.

$$A_1(f) = \begin{cases} |1 - g\gamma|, & \text{when } f = f_1, \\ |\alpha - g|, & \text{when } f = f_2, \\ |(1+\alpha) - g(1+\gamma)|, & \text{when } f = f_3, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

$$A_2(f) = \begin{cases} |\gamma - g|, & \text{when } f = f_1, \\ |1 - g\alpha|, & \text{when } f = f_2, \\ |(1+\gamma) - g(1+\alpha)|, & \text{when } f = f_3, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The maximum value that g can assume is 1. The frequency-azimuth plane exhibits nulls at the frequencies f_1, f_2 and f_3 for the following gains,

$$A(f) = 0 \text{ if } \begin{cases} g = \gamma, & \text{when } f = f_1, \\ g = \alpha, & \text{when } f = f_2, \\ g = \frac{1+\alpha}{1+\gamma} \text{ or } \frac{1+\gamma}{1+\alpha} \text{s.t.} g \leq 1 & \text{when } f = f_3. \end{cases} \tag{13}$$

We can now summarize the operation of the Adress algorithm and introduce the challenge addressed by this paper. Adress computes windowed Discrete Fourier Transforms, $X_1(k)$ and $X_2(k)$, of the left and right channel mixture signals above, which are discrete time signals, and have been sampled at a sufficiently high sampling rate. This produces M frequency bins, $k = 0, 1, \ldots, M$. The frequency-azimuth matrix is constructed by forming the matrix, $\mathbf{A} \in \mathbb{R}^{M \times N}$, which examines the two-channel mixture in each frequency bin, k, using a set of values of gain values, g, in the range $g = \left\{\frac{0}{\beta}, \frac{1}{\beta}, \ldots \frac{\beta}{\beta}\right\}$, where $\beta = \frac{N}{2}$. Reconstruction of the component sources is achieved by assigning TF bins to sources depending on the location of the nulls in the frequency-azimuth plane.

B. Problem

The frequency-azimuth plane produces nulls at $g = \alpha$ which corresponds to the source $s_2(t)$ in the mixture for frequency f_2. However, for the f_3 frequency component of this source the null is generally not located at α. Instead it is located at $\frac{1+\alpha}{1+\gamma}$ or $\frac{1+\gamma}{1+\alpha}$ depending on the size of α and γ. Similarly, the frequency-azimuth plane produces nulls at $g = \gamma$ which corresponds to the source $s_1(t)$ in the mixture for the frequency component f_1, but the null is generally not located at this scale factor for f_3. Co-occupation in the f_3 frequency causes the Adress algorithm to assign all of the energy of the frequency f_3 to one of the sources and none of the energy to the other sources in the mixture. This assignment is done based on the distance of the null of the f_3 component from the nulls for the rest of the s_1 and s_2 frequency-gain nulls. The absence of the f_3 frequency component in one of the signals causes what is sometimes called musical noise to appear in the reconstructed source signals. In this example 50% of the frequency components of one of the sources will be missing and the other source will have a magnitude which is too large for that missing frequency component. This type of problem has been identified and called Frequency-Azimuth Smearing in the literature but it has not been solved.

Methods for measuring the level of disjointness of two sources in the TF domain are called WDO [2], [12]. They have been used to determine what parametrization of the Short-Time Fourier Transform (STFT) will give the most disjoint, or non-overlapping representation of the source signals in the TF mixtures. We now present a solution to the problem which is motivated by re-considering the Adress mixing model. Our

argument is presented in the context of the simple case above. We then demonstrate it is applicable in the more general case of arbitrary mixtures.

III. Redress: De-mixing Co-occupied TF bins

The Adress algorithm finds nulls in the frequency-azimuth plane and then assigns all frequency bins, or bands, which have nulls in similar positions to the same source signal. We now introduce how the Redress algorithm considers the full range of gain values as opposed to just the location of the nulls. Redress attempts to re-express the entire frequency-azimuth plane as a factorization in order to estimate how much each source is contributing to each frequency. This is important when frequencies are occupied by more than one of the input source signals. The motivation for Redress is that the entire set of gains that are used to examine each TF bin have information; this information should be used to help de-mix TF bins which are shared between multiple sources.

To explain how this is done we consider the construction of the frequency-azimuth plane $A(f)$, starting with the $A_1(f)$ component. Due to the symmetry of the magnitude TF plane we only examine the positive frequencies. The first component $A_1(f)$ may be expressed as the difference between the Fourier transforms of the mixtures, where the channel, $X_2(f)$, is scaled successively by g in order to find nulls,

$$|X_1(f) - gX_2(f)| = |(1-g\gamma)\delta(f-f_1) + (\alpha-g)\delta(f-f_2) + ((1+\alpha)-g(1+\gamma))\delta(f-f_3)|. \quad (14)$$

This difference can be re-expressed more compactly as

$$|X_1(f) - gX_2(f)| = |(1-g\gamma)S_1(f) + (\alpha-g)S_2(f)|. \quad (15)$$

Similarly, for the second component we obtain,

$$|X_2(f) - gX_1(f)| = |(\gamma-g)S_1(f) + (1-g\alpha)S_2(f)|. \quad (16)$$

Redress Magnitude TF mixing model: Working with discrete TF mixtures, the frequency-azimuth plane may be expressed in matrix form, $\mathbf{A}$, as the product of a set of source frequency basis functions and source azimuth activation functions,

$$\mathbf{A} \approx \mathbf{WH}, \quad (17)$$

where $\mathbf{W} \in \mathbb{R}_+^{M\times R}$ is a set of source frequency basis functions and $\mathbf{H} \in \mathbb{R}_+^{R\times N}$ is a set of source azimuth activation functions. In short, the frequency-azimuth plane is approximated by $R = 2$ frequency basis functions above. One of these basis continuous frequency domain functions should have energy at $f = f_1$ and $f = f_3$ in order to represent $s_1(t)$. The second basis function should have energy at $f = f_2$ and $f = f_3$ to represent the second source, $s_2(t)$. In the example used to motivate this argument above the source discrete frequency basis functions are $\mathbf{W}_{:,1} = |S_1(k)|$ and $\mathbf{W}_{:,2} = |S_2(k)|$, subject to permutation ambiguity. The source azimuth activation functions are

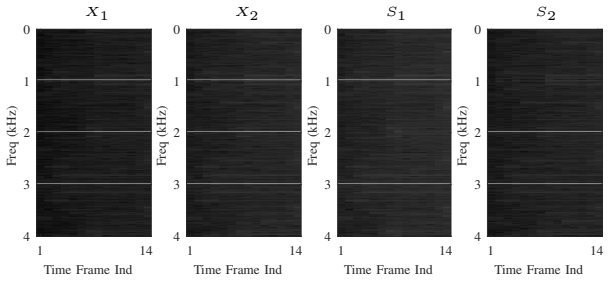

Fig. 1. The positive frequencies of the magnitude spectra of the mixtures, $X_1(k)$ and $X_2(k)$, and source signals, $S_1(k)$ and $S_2(k)$, are illustrated from left to right. The frequency components in $s_1(t)$ are $f_1 = 1$kHz and $f_3 = 3$kHz respectively. The frequency components in $s_2(t)$ are $f_1 = 2$kHz and $f_3 = 3$kHz respectively.

$$H_{1,1:\frac{M}{2}} = [|(1-g\gamma)|]_{0\le g\le 1} \quad (18)$$

$$H_{2,1:\frac{M}{2}} = [|\alpha-g|]_{0\le g\le 1} \quad (19)$$

$$H_{1,\frac{M}{2}+1:M} = [|(\gamma-g)|]_{0\le g\le 1} \quad (20)$$

$$H_{2,\frac{M}{2}+1:M} = [|1-g\alpha|]_{0\le g\le 1} \quad (21)$$

We justify re-writing the absolute value of the difference in $|X_1(f) - gX_2(f)|$ as the difference of the absolute value of the terms as follows. There are two cases to consider. In the first case, a frequency bin is only occupied by one source. When $f = f_1$, then $|X_1(f_1) - gX_2(f_1)| = |(1-g)S_1(f_1)|$ which can be written as $|X_1(f_1) - gX_2(f_1)| = |(1-g)||S_1(f_1)|$. The matrix approximation in Eqn. 17 is an accurate approximation in this case. In the second case, a frequency bin is occupied by two sources. $|X_1(f_3) - gX_2(f_3)| = |((1+\alpha) - g(1+\gamma))\delta(f-f_3)|$. In general the component $\delta(f-f_3)$ will be complex-valued; however, as the mixing model does not consider relative delays between channels this complex-valued variable will be the same for both channels and thus we can re-write $|X_1(f_3) - gX_2(f_3)| = |((1+\alpha)-g(1+\gamma))||\delta(f-f_3)|$. Because $\alpha > 0$, and $\gamma > 0$ we can express this difference as $|X_1(f_3) - gX_2(f_3)| \approx (|(1+\alpha)| - |g(1+\gamma)|)|\delta(f-f_3)|$, by appealing to the triangle inequality.

IV. Reconstruction

Basis-Activation Factorization: We now describe how to learn this basis function-activation factorization and to recover the source signals. Given the frequency-azimuth matrix $\mathbf{A}$, which has been generated from the discrete frequency representations of the signals, we minimize the Frobenius norm between the frequency-azimuth matrix $\mathbf{A}$ and the current estimate of this matrix, $\mathbf{WH}$, which is denoted $D_F(\mathbf{A}||\mathbf{WH})$ and defined as

$$D_F(\mathbf{A}||\mathbf{WH}) = \frac{1}{2}\sum_{ik} |a_{ik} - [\mathbf{WH}]_{ik}|^2. \quad (22)$$

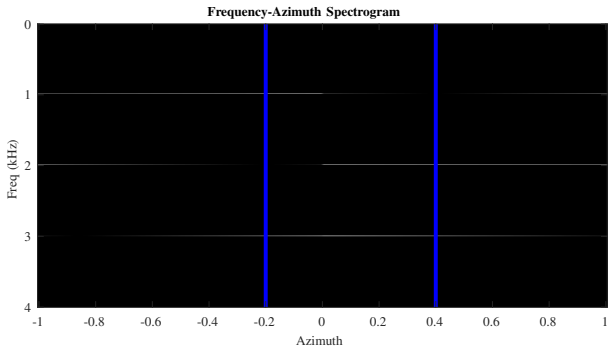

Fig. 2. The intensity (white for high and black for low) of the azimugram is illustrated. Black regions in horizontal white lines indicate nulls in the response. The null for f_1 is at .4 which indicates the pan-position of $s_1(t)$ [vertical blue line]. The null for f_2 is at $-.2$, (e.g. $-\alpha$) indicating the pan-position of $s_2(t)$ [vertical blue line].

A suitable step-size parameter was introduced in [10] which resulted in an alternating, multiplicative, gradient descent updating algorithm comprising of the two update rules

$$\mathbf{W} \leftarrow \mathbf{W} \odot \mathbf{A}\mathbf{H}^T \oslash \mathbf{W}\mathbf{H}\mathbf{H}^T, \quad (23)$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \mathbf{W}^T\mathbf{A} \oslash \mathbf{W}^T\mathbf{W}\mathbf{H}, \quad (24)$$

where $\odot$ represents element-wise multiplication and $\oslash$ represents element-wise division. The authors argue that the advantage of having multiplicative updates is that it provides a simple update rule under which $\mathbf{W}$ and $\mathbf{H}$ never become negative and so projection into the positive orthant of the space is not necessary. The veracity of this statement is borne out by the success of the algorithm and the range of extensions of the technique [9], [13]. The alternating nature of the updates implies that the optimization in no longer convex. As NMF is not convex when the update rules are alternated the solution is not guaranteed to be unique or exact [10]. The success of the approach is evaluated for the problem introduced and then investigated for more general mixtures.

Source Recovery: For each azimuth-frequency matrix, $\mathbf{A}$, which is generated for each position of the analysis window, the columns of $\mathbf{W}$ may correspond to different sources each time the decomposition is computed [9], [10], [13]. Reconstruction of the magnitude spectrograms of the source signals relies on the ability to address this permutation ambiguity. To undo this permutation, the rows of the source azimuth activation matrix $\mathbf{H}$ are searched in order to find the locations of the minima in each row. Rows with similar minima locations, which correspond to the gains used to examine the mixtures, are assigned to the same source. An approximation of each source's contribution to each windowed position of the mixtures, either $X_1(k)$ or $X_2(k)$, is determined by minimizing the Frobenius norm of the difference between one of the mixtures, for example $\mathbf{V} = \mathbf{X}_1$, and a low-rank approximation of this matrix, e.g. $\mathbf{WH}$. This minimization is done with respect to $\mathbf{H}$ as the basis functions for each source were learned in the previous step and with respect to each mixture.

$$\min_{\mathbf{H}} \sum_{ik} |v_{ik} - [\mathbf{W}\mathbf{H}_{ik}]|^2 \text{ given } \mathbf{W}. \quad (25)$$

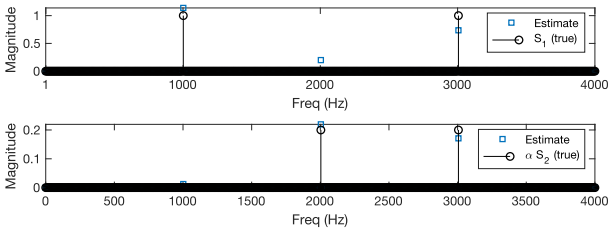

Fig. 3. Columns of the magnitude spectra of the sources and the two frequency basis functions, $\mathbf{W}$, are illustrated. The functions, $\mathbf{W}$, both contain the f_3 frequency component. Adress does not allow both of the reconstructed sources to contain this component.

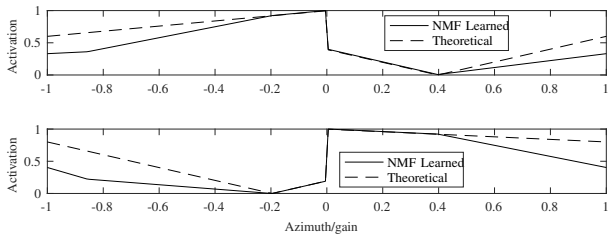

Fig. 4. Theoretical and estimated azimuth activations, $\mathbf{H}$, for both sources. The azimuth activation minima lie at the correct pan-positions.

It is achieved by running the $\mathbf{H}$-update in Eqn. 23 in an iteration and holding the $\mathbf{W}$ matrix fixed. Each rank-1 approximation is an estimate of the magnitude spectrum of a source. In order to produce discrete time-domain sources we use the phase of one of the mixture signals and apply an inverse short-time Fourier transform. For example, we approximate the first source, $\hat{S}_1$, using the rank-1 approximation of the mixture, $\mathbf{W}_{:,1}\mathbf{H}_{1,:}$ and the first mixture's phase,

$$\hat{S}_1 = (\mathbf{W}_{:,1}\mathbf{H}_{1,:}) \odot \exp(i \arg(X_1)). \quad (26)$$

V. Experiments

We demonstrate the effects of overlapping sources in the TF domain on the reconstruction achieved by Adress and on the Redress algorithm on the audio mixtures presented above. We then compare the performance of Adress and Redress on real mixtures of two instruments and three instruments.

A. Redress Applied to Overlapping Audio

To evaluate the performance of Redress, we apply it to a music example. The source, $s_1(t)$, consists of two sinusoids with frequencies 1kHz and 3kHz. The source $s_2(t)$ consists of two sinusoids with frequencies 2kHz and 3kHz. These sources are rudimentary musical sounds, where each synthesised instrument/source does not produce harmonics. The amplitudes of all sinusoids are 2V. The pan-mixing weights used to create the observed mixtures, $x_1(t)$ and $x_2(t)$, are $\alpha = .2$ and $\gamma = .4$. Both Adress and Redress are applied to the pan-mixed signals and the recovered sources are presented. Fig. 1 illustrates the magnitude spectra of the mixture signals and the source signals. The sampling rate is 8kHz. The mixtures examined

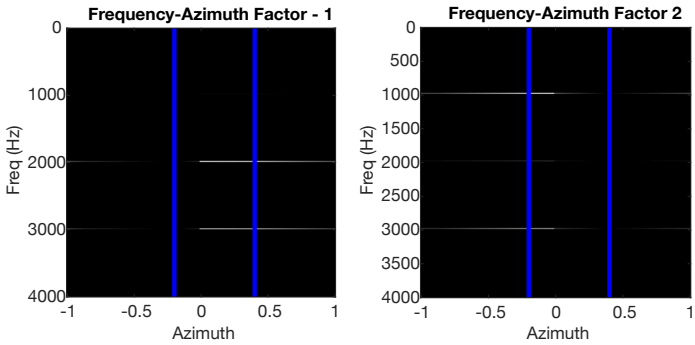

Fig. 5. Rank-1 components of the frequency-azimuth plane are illustrated. One rank-1 component has its null at the same gain for 1kHz and 3kHz. One rank-1 component has its null at the same gain for 2kHz and 3kHz. Both reconstructed sources are assigned some portion of the energy at 3kHz.

TABLE I
ROUNDED SNR OF INSTRUMENT ESTIMATES FOR TWO SOURCES USING REDRESS AND ADRESS.

	s_1, s_2 Piano, Bass	s_1, s_2 Piano, Strings	s_1, s_2 Bass, Strings
Redress	19.2dB, 17.9dB	11.9dB, 12.0dB	12.1dB, 9.6dB
Adress	10.6dB, 12.4dB	4.3dB, 4.2dB	6.6dB, 6.0dB

TABLE II
ROUNDED SNR OF INSTRUMENT ESTIMATES FOR THREE SOURCES USING REDRESS AND ADRESS.

	s_1, s_2, s_3 Piano, Bass, Strings	s_1, s_2, s_3 Bass, Piano, Pads
Redress	3.1dB, 2.3dB, 1.4dB	5.2dB, 3.5dB, 5.2dB
Adress	3.7dB, 5.6dB .9dB	6dB 6.5dB, 7.7dB

are 1s long. The analysis window used is .125s long and it is advanced by 0.0625s. The Redress algorithm decomposes the frequency-azimuth matrix with a rank-2 approximation, $R = 2$, using 100 iterations. The mixtures are examined using $\beta = 100$. Fig. 2 illustrates the resulting frequency-azimuth plane. Negative values on the x-axis indicate the left channel, which has been flipped left-right in order to preserve the symmetry of the gain values, where an azimuth of zero corresponds to a gain, $g = 0$, and a negative azimuth corresponds to a positive gain on the left channel. The nulls for the f_1 and f_2 components indicate the pan-positions of the two sources, $s_1(t)$, has a null at $\gamma = .4$ and $s_2(t)$ has a null at $\alpha = -.2$. Vertical blue lines are superimposed to illustrate these locations. Crucially the null for the f_3 component is not at the same pan-position as either of the two sources. Adress assigns the f_3 component to one of the sources, $s_2(t)$, but not $s_1(t)$, resulting in a missing frequency, f_3, in the reconstructed spectrum for $s_1(t)$.

Fig. 3 illustrates the true magnitude spectra (at one time index) for the sources and the estimated magnitude spectra, from $\mathbf{W}$, using Redress. The estimate for source s_1 has non-zero energy at f_2Hz which indicates that this source component has not been completely de-mixed. In addition the second source does not have the appropriate energy for the f_3Hz frequency. A significant advantage of the Redress approach over the Adress algorithm is that both estimates of the sources have energy in the f_3Hz frequency. This demonstrates its ability to de-mix, mixed TF bins.

We plot the theoretical azimuth activation functions and overlay the estimated azimuth activation functions, $\mathbf{H}$, in Fig. 4. The estimates give a good approximation of the location of the nulls in the two activation functions. The shape of the estimated functions follow the theoretical functions closely. Note that at the location of the null in the f_3 component in the mixture frequency azimuth plot in Fig. 2, e.g. $\frac{1+\alpha}{1+\gamma} = .857$, that there is a sharp change in the slope of the two estimated activation functions. Although this component has been de-mixed, the de-mixture achieved is an estimate. Finally, we plot rank-1 approximations of the frequency-azimuth plot in Fig. 5 and observe that both rank-1 estimates have frequency components at f_3Hz. The nulls observed for both sources for the f_3 component are at the same position as the f_1 and f_2 frequency components respectively. We now consider de-mixing pan-mixed instruments.

B. Redress Applied on Music

In order to evaluate Redress in a realistic scenario we downloaded the component instrument tracks for a song, *Only Love* by Shannon Hurley. These were made available here (http://ccmixter.org/shannon-hurley) in order to solicit remixes from the public. The instruments were sampled at 44.1kHz. We used a 4096-sample Hamming window which is advanced by 2048 samples in our analysis. The rank of the decomposition used was $R = 2$ for the two-instrument mixes and $R = 3$ for the three instrument mixes, and the decomposition iteration was run for 100 steps. We set $\beta = 100$. For Adress, we used an azimuth subspace width, $H = 20$, [1]. Each track was $\approx$ 330s long, however, in the song some instruments were only played during certain intervals and so we evaluated Redress using intervals where both of the instruments in the mix were playing. In many recordings a number of instruments are positioned at the same pan-position [1] and so we considered mixtures consisting of instruments in two or three different pan-positions. The Strings *instrument* for example was a String section which consisted of a double bass, violins, violas, etc. In the first case we pan-mixed two instrument mixes consisting of: (1) Piano and Bass; (2) Piano and Strings; and finally (3) Bass and Strings. The average Signal-to-Noise Ratio of the source estimates over a range of pan-positions are summarized in Table I.

The Piano and Strings sources exhibited significant overlap in the TF domain due to the number of instruments present in the String section. We posit that this caused the reconstruction SNR to be reduced for Redress and Adress. De-mixing achieved an SNR of approximately 18dB and 19dB for mixtures consisting of the Piano and the Bass for Redress. This was due to the relatively little overlap between TF representation of the instruments. Adress however achieved a lower SNR for both instruments for the same mixtures. Although separation of the Strings and the Bass was relatively successful –frequencies corresponding to the Bass source

were generally extracted– all of the string component sources (violin, viola, etc.) in the mixture were sometimes not fully separated-out from the Bass source estimate which caused there to be significant overlap between the Bass and the bass component of the strings section. This caused the SNR achieved by Redress to be reduced; this SNR was still higher than the SNR achieved by Adress. Similarly there was a high degree of overlap between the Strings and the Piano instrument which Redress did not fully separate. Redress did outperform Adress with respect to SNR for this mixture.

In the second case we pan-mixed three instruments consisting of: (1) Piano, Bass and Strings and (2) Bass, Piano and Pads. As expected there was a decrease in the SNR of the reconstructed sources as the overlap of the TF content of the sources increased in Table II for both Adress and Redress. The overlap of the bass component of the Strings section and the Bass –which caused a decrease in the SNR of the de-mixtures in the two source case– was once again evident in the three source mixture. When both of these sources were present the TF overlap was higher and thus the SNR of both of these sources was lower than for the other source component. In the Bass-Piano-Pads mixture case, the reconstruction SNR of the sources was high because the source signals did not overlap as frequently in the TF domain. In the three source mixtures, the Adress algorithm outperformed the Redress approach in terms of SNR by approximately $1-2$dB; however in the two source case, Redress achieved an SNR which was a factor of two better than the SNRs achieved by Adress for some instruments. We conclude that Redress helps to de-mix TF bins where multiple sources are present if in general this has a low frequency of occurring in the TF representation. However, if there is a high degree of TF overlap, the type of hard-masking approach adopted by Adress yields better SNRs. This result is consistent with previous analysis of the DUET algorithm [2], [3]. In future work we will reconsider the reconstruction approach used by Redress. One direction of improvement lies in allowing Redress to adapt the factorization rank, R, in response to the changing amplitudes of source signals, in particular the Bass the examples above.

In order to give an indication of the perceptual performance of Redress we illustrate the true piano and bass sources and the estimates of these sources learned by Redress in Fig. 6. In addition, listening tests of the recovered instruments reveal that in many cases the separation achieved was good; the high SNRs achieved in Table I support this claim.

VI. Conclusions

The Adress algorithm, which uses gain scaling techniques to expose frequency dependent nulls across the azimuth plane, struggles when two or more sources occupy the same TF bin. We have proposed a method, namely Redress, to de-mix the contribution of multiple sources to a TF bin. We believe that performance improvement could be achieved by Redress by reconsidering how reconstruction is performed. When there is only one source playing, the Redress approach attempts to learn basis functions for R source signals even though there is only one source present. For example in the piano-bass mix considered, when the piano is a more dominant signal in the TF domain than the bass, some of the piano signal is de-mixed into the bass. This does not occur at the onset of a bass note, when the bass dominates, but when the amplitude of the bass has begun to decrease. Solving this problem, by incorporating Adress as a pre-processing step may improve Redress reconstruction in the three source case.

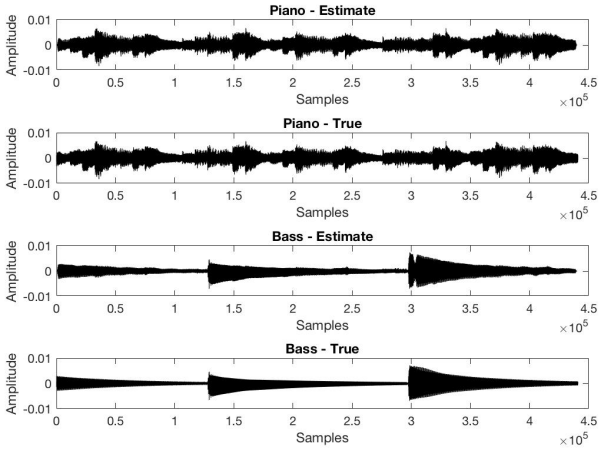

Fig. 6. Piano and bass: true source signals and Redress estimates. The piano signal estimate is accurate but the bass signal estimate still has some of the piano signal mixed in with it.

References

[1] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *7th DAFX*, 2004.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[3] R. de Fréin and S. T. Rickard, "Power-weighted divergences for relative attenuation and delay estimation," *IEEE Sig. Proc. Let.*, vol. 23, no. 11, pp. 1612–1616, Nov 2016.

[4] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *IEEE WASPAA*, 2003, pp. 55–58.

[5] T. Melia, S. Rickard, and C. Fearon, "Histogram-based blind source separation of more sources than sensors using a DUET-ESPRIT technique," in *13th EUSIPCO*, 2005, pp. 1–4.

[6] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ilrma originating fromICA and NMF," *APSIPA Trans. on Sig. Inf. Proc.*, vol. 8, 2019.

[7] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environments using DESPRIT," *EURASIP J. Adv. Sig. Proc.*, 2007.

[8] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation of many speech signals using near-field and far-field models," *EURASIP J. Adv. Sig. Proc.*, vol. 2006, p. 200, Jan. 2006.

[9] P. D. O'Grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *16th IEEE MLSP*, 2006, pp. 427–432.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*. MIT Press, 2000, pp. 556–562.

[11] R. de Fréin, "Learning and storing the parts of objects: IMF," in *IEEE Int. Wrkshp MLSP*, 2014, pp. 1–6.

[12] R. de Fréin and S. T. Rickard, "The Synchronized Short-Time-Fourier-Transform: Properties and definitions for multichannel source separation," *IEEE Trans. Sig. Proc.*, vol. 59, no. 1, pp. 91–103, Jan 2011.

[13] ——, "Learning speech features in the presence of noise: Sparse convolutive robust non-negative matrix factorization," in *16th Int. Conf. DSP*, July 2009, pp. 1–6.