Audio Engineering Society

Convention Paper

Presented at the 117th Convention
2004 October 28–31 San Francisco, CA, USA

Real-time Sound Source Separation: Azimuth Discrimination and Resynthesis

Dan Barry[1], Bob Lawlor[2], and Eugene Coyle[3]

[1] Dept. of Control Systems and Electrical Engineering, Dublin Institute of Technology, Kevin St, Dublin 8, Ireland.
barrydn@eircom.net

[2] Dept. of Electronic Engineering, National University of Ireland, Maynooth, Ireland
rlawlor@eeng.may.ie

[3] Dept. of Control Systems and Electrical Engineering, Dublin Institute of Technology, Kevin St, Dublin 8, Ireland.
eugene.coyle@dit.ie

ABSTRACT

We present a real-time sound source separation algorithm which performs the task of source separation based on the lateral displacement of a source within the stereo field. The algorithm exploits the use of the pan pot as a means to achieve image localisation within stereophonic recordings. As such, only an interaural intensity difference exists between left and right channels for a single source. Gain scaling and phase cancellation techniques are used in the frequency domain to expose frequency dependent nulls across the azimuth plane. The position of these nulls in conjunction with magnitude estimation and grouping techniques are then used to resynthesise separated sources. Results obtained from real recordings show that for music, this algorithm outperforms current source separation schemes.

1. INTRODUCTION

Our research is concerned with extracting sound sources from stereo music recordings for the purposes of audition and analysis. This is termed sound source separation and has been the topic of extensive research in recent years. In general, the task is to extract individual sound sources from some number of source mixtures. Many of the current approaches to this problem fall into one of two main categories, Independent Component Analysis, (ICA) [1],[2] and Computational Auditory Scene Analysis, (CASA) [3]. ICA is a statistical source separation method which operates under the assumption that the latent sources have the property of mutual statistical independence and are non-gaussian. In addition to this, ICA assumes that there are at least as many observation mixtures as there are independent sources. Since we are concerned with musical recordings, we will have at most only 2

observation mixtures, the left and right channels. This makes pure ICA unsuitable for the problem where more than two sources exist. One solution to the case where sources out number mixtures is the DUET algorithm [4], [5]. In order for this method to be successful, the independent sources must be approximately 'W-disjoint orthogonal'. This condition effectively means that each of the latent sources do not overlap significantly in the time or the frequency domain. It was shown that speech does indeed approximate this condition and so the method is applicable to the case where the mixture signals contain only speech. It should be appreciated that western tonal music will contain a significant amount of overlap in both the time and frequency domain and so the method fails for such music mixtures. CASA methods on the other hand, attempt to decompose a sound mixture into auditory events which are then grouped according to perceptually motivated heuristics [6], such as common onset and offset of harmonically related components, or frequency and amplitude co-modulation of components. We present a novel approach which we term Azimuth Discrimination and Resynthesis (ADRess). The approach we describe is a fast and efficient way to perform sound source separation on the majority of stereophonic recordings.

2. BACKGROUND

Since the advent of multi-channel recording systems in the early 1960's, most musical recordings are made in such a fashion whereby *N* sources are recorded individually, then electrically summed and distributed across 2 channels using a mixing console. Image localisation, referring to the apparent position of a particular instrument in the stereo field, is achieved by using a panoramic potentiometer. This device allows a single sound source to be divided into to two channels with continuously variable intensity ratios [7]. By virtue of this, a single source may be virtually positioned at any point between the speakers. So localisation is achieved by creating an interaural intensity difference, (IID). This is a well known phenomenon [8]. The pan pot was devised to simulate IID's by attenuating the source signal fed to one reproduction channel, causing it to be localised more in the opposite channel. This means that for any single source in such a recording, the phase of a source is coherent between left and right, and only its intensity differs. It is precisely this that allows us to perform our separation. This mixing model is also assumed in [9],[10]and[11]. It must be noted then, that our method is only applicable to recordings such as described above. Binaural or Stereo Pair recordings will not respond as well to this method although we have had some success in these cases also. Theoretically, the method should work for the Mid-Side technique under non reverberant conditions since the apparent position of a source is encoded as an intensity difference using this method.

3. METHOD

Firstly we obtain an STFT of each channel. Then, on a frame by frame basis, gain scaling is applied to one of the channels so that one source's intensity becomes equal in both left and right channels. A simple subtraction of the channels will cause that source to cancel out due to phase cancellation. The cancelled source is then recovered by first creating a frequency-azimuth plane *(section 3.1)* which is analyzed for local minima along the azimuth axis. These local minima represent points at which some gain scalar caused phase cancellation. It is observed that at some point where a source cancels out, only the frequencies which were present in the source will show a local minimum. These minima signify energy loss due to source cancellation. It is shown that this energy loss is proportional to the amount of energy which the cancelled source had contributed to the overall mixture. The magnitudes of these minima are then estimated and assigned a phase after which an IFFT in conjunction with an overlap add scheme is used to resynthesise the cancelled source.

3.1. Azimuth Discrimination

The mixing process we have described can be expressed as,

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

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

where S_j are the J independent sources, Pl_j and Pr_j are the left and right panning co-efficients for the j^{th} source, and L and R are the resultant left and right channel mixtures. Our algorithm takes $L(t)$ and $R(t)$ as it's inputs and attempts to recover S_j, the sources. We can see from equation 1a and 1b that the intensity ratio of the j^{th} source, $g(j)$, between the left and right channels can be expressed as,

$$g(j) = Plj / Prj \tag{2}$$

This implies that $Pl_j=g(j).Pr_j$. So, multiplying the right channel R by $g(j)$ will make the intensity of the j^{th} source equal in left and right. And since L and R are simply the superposition of the scaled sources, then $L - g(j).R$ will cause the j^{th} source to cancel out. In practice we use $L - g(j).R$, if the j^{th} source is predominant in the right channel and $R - g(j).L$ if the j^{th} source is predominant in the left channel. This serves two purposes, firstly it gives us a range for $g(j)$ such that: $0 \leq g(j) \leq 1$. Secondly, it insures that we are always scaling one channel down in order to match the intensities of a particular source, thus avoiding infinitely large scaling factors.

So far we have only described how it is possible to cancel a source assuming the mixing model we have presented. Next we will deal with recovering the cancelled source. In order to do this we must move into the frequency domain. We divide the stereo mixture into short time frames and carry out an FFT on each:

$$Lf(k) = \sum_{n=0}^{N-1} L(n) W_n^{kn} \quad (3a)$$

$$Rf(k) = \sum_{n=0}^{N-1} R(n) W_n^{kn} \quad (3b)$$

where $Wn=e^{-j2\pi/N}$ and Lf and Rf are short time frequency domain representations of the left and right channels respectively. In practice we use a 4096 point FFT with a Hanning window and an analysis step size of 1024 points. We create a frequency-azimuth plane for left and right channels individually, see figure *2*. The azimuth resolution, β, refers to how many equally spaced gain scaling values of g we will use to construct our frequency-azimuth plane. We relate g and β as follows,

$$g(i) = i.(1/\beta) \quad (4)$$

for all i where, $0 \leq i \leq \beta$, and where i and β are integer values.

Large values of β will lead to more accurate azimuth discrimination but will increase the computational load. Assuming an N point FFT, our frequency-azimuth plane will be an N x β array for each channel. The right and left frequency-azimuth plane are then constructed using equations 5a and 5b,

$$Az_R(k,i) = | Lf(k) - g(i).Rf(k) | \quad (5a)$$

$$Az_L(k,i) = | Rf(k) - g(i).Lf(k) | \quad (5b)$$

for all i and k where, $0 \leq i \leq \beta$, and $1 \leq k \leq N$.

It must be stated that we are using the term "azimuth" loosely. We are not dealing with angles of incidence. The azimuth we speak of is purely a function of the intensity ratio, created by the pan pot during mix down. In order to illustrate how this process reveals frequency dependent nulls, we generated two test signals, each with 5 unique partials. A stereo mix was created such that both sources were panned to the right, but each with a different intensity ratio. Using this test signal, the frequency-azimuth plane in figure 1 was created using equation 5a, with, $\beta=100$, and N=1024 point FFT. It can clearly be seen that partials from each source are at a minimum at the same point along the azimuth axis as in figure 1 and figure 2

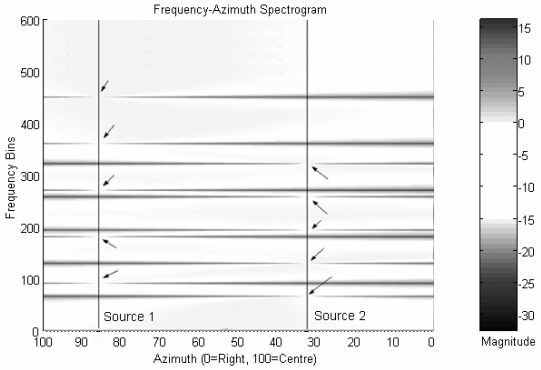

Figure 1: The Frequency-Azimuth spectrogram for the right channel. We used 2 synthetic sources each comprising of 5 non-overlapping partials. The arrows indicate frequency dependent nulls caused by phase cancelation.

In order to estimate the magnitude of these nulls we redefine equations 5a and 5b as 6a and 6b:

$$Az_{R(k,i)} = \begin{cases} Az_{R(k)_{max}} - Az_{R(k)_{min}}, & \text{if } Az_{R(k,i)} = Az_{R(k)_{min}} \\ 0, & \text{otherwise.} \end{cases} \quad (6a)$$

$$Az_{L(k,i)} = \begin{cases} Az_{L(K)_{max}} - Az_{L(K)_{min}}, & \text{if } Az_{L(k,i)} = Az_{L(k)_{min}} \\ 0, & \text{otherwise.} \end{cases} \quad (6b)$$

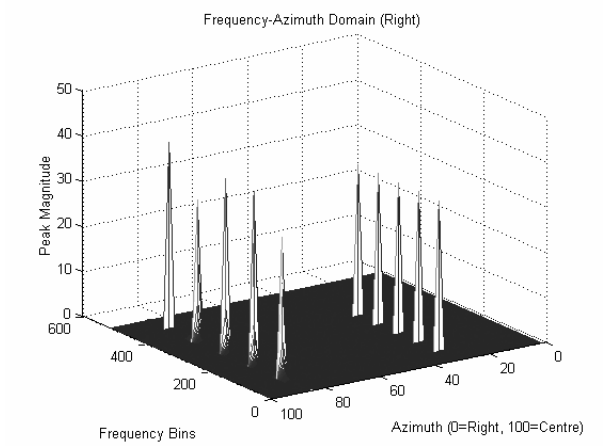

Figure 2: The Frequency-Azimuth plane for the right channel. The magnitudes of the frequency dependent nulls have been estimated. The harmonic structure of each source is now clearly visible as is their spatial distribution.

Effectively, we are turning nulls into peaks as can be seen in figure 2. However, the test signal described, represents the ideal case where there is no harmonic overlap between 2 sources. This is almost never the case when it comes to tonal music. Harmony is one of the fundamentals of music creation, and as such instruments will more often than not be playing harmonically related notes simultaneously which implies that there will be significant harmonic overlap with real musical signals. The result of this is that frequencies will not group themselves as neatly across the azimuth plane as in figure 2. We have observed "frequency-azimuth smearing". This is caused when two or more sources contain energy in a single frequency bin. The apparent frequency dependent null drifts away from a source position and may be at a minimum at a position where there is no source at all. For instance, if two sources in different positions, contained energy at a particular frequency, the apparent null will appear somewhere between the two sources. To over come this problem, we define an "azimuth subspace width", H, such that $1 \leq H \leq \beta$. This allows us to recover peaks within a given neighborhood. These azimuth subspaces may overlap and often do. Nulls that drift away from their source positions can now be re-included for resynthesis. A wide azimuth subspace will result in worse rejection of nearby sources. On the other hand a narrow azimuth subspace will lead to poor resynthesis and missing frequency information. This parameter is varied depending on source proximity. Figure 3 shows the same two test signals as before only each includes one extra partial of the same frequency. It can clearly be seen that the common partial is now apparent between the two sources. In order to recover it, the azimuth subspace boundary of the source must extend beyond it. This is shown for source one. At this point we introduce the "discrimination index", d where, $0 \leq d \leq \beta$. This index, d, along with the azimuth subspace width, H, will define what portion of the frequency-azimuth plane is extracted for resynthesis.

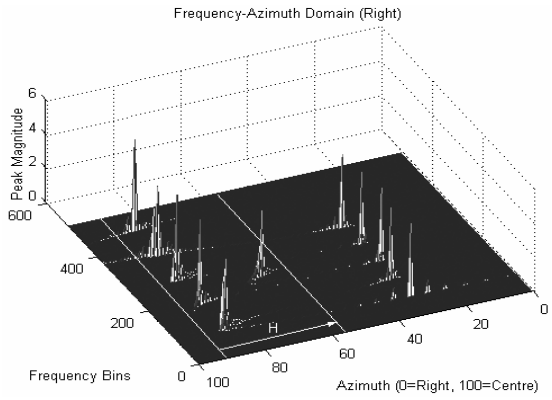

Figure 3: The Frequency-Azimuth Plane. The common partial is apparent between the 2 sources. The azimuth subspace width for source 1, H, is set to include the common partial.

3.2 Source Resynthesis

In order to resynthesise only one source, we set the discrimination index, d, to the apparent position of the source. In figure 3, there are 2 sources, one at approximately 85 points along the azimuth axis, and the other at 33. The azimuth subspace width, H, is then set such that the best percieved resynthesis quality is achieved. In practice, we centre the azimuth subspace over the discrimination index such that the subspace spans from $d-H/2$ to $d+H/2$. The peaks for resyntheis are then extracted using equations 7a and 7b,

$$Y_{R(k)} = \sum_{i=d-H/2}^{i=d+H/2} Az_{R(k,i)} \quad 1 \leq k \leq N \qquad (7a)$$

$$Y_{L(k)} = \sum_{i=d-H/2}^{i=d+H/2} Az_{L(k,i)} \quad 1 \leq k \leq N \qquad (7b)$$

The resultant Y_R and Y_L are $1 \times N$ arrays containing only the bin magnitudes pertaining to a particular azimuth subspace as defined by d and H. More specifically, Y_R

and Y_L contain the short time power spectrum of the separated source. At this point it should be noted that, if two sources have the same intensity ratio, i.e. they share the same pan position, both will be present in the extracted subspace. This is particularly true of the "centre" position. It is common practice in audio mix down to place a number of instruments here, usually voice and very often bass guitar and elements of the drum kit too. In this instance, band limiting can be used to further isolate the source of interest. Noise reduction is sometimes favorable depending on the characteristics of the audio for separation. This is crudely implemented by applying a threshold to the magnitude spectra before resynthesis.

The bin phases could be estimated using a technique such as 'magnitude only reconstruction' but we have found that using the original bin phases is adequate, equation 8a and 8b. Once we have bin phases and magnitudes we can convert from polar to complex form after which the azimuth subspace is resynthesised using the IFFT, equation 9.

$$\Phi_{R(k)} = \angle(Rf_{(k)}) \tag{8a}$$

$$\Phi_{L(k)} = \angle(Lf_{(k)}) \tag{8b}$$

We resynthesise our short time signal using the IFFT,

$$X_{(n)} = \frac{1}{N}\sum_{k=1}^{N} X_{(k)} W_n^{-kn} \tag{9}$$

where $W_n = e^{-j2\pi/N}$

The resynthesised time frames are then recombined using a standard overlap and add scheme.

4. REAL-TIME IMPLEMENTATION

A real-time version of the algorithm has been implemented. The interface contains 3 user controls:

1. Azimuth Index; which is represented by the parameter *d* above and is responsible for selecting the position or source for separation.
2. Azimuth Range; which is represented by the parameter *H* above and controls how wide or narrow the separation subspace will be.
3. Noise Threshold; which is a basic form of noise reduction.

The azimuth resolution, β, is fixed at a value of 10 for each channel yielding 20 discrete azimuth points in the stereo field. This was also necessary to ensure real-time operation of the algorithm. The interface also contains an animated graphical display containing real-time information about the distribution of sources across the stereo field. The graph illustrates the presence of a source with a peak. The user can then use both audio and visual feedback in order to set the parameters for the best percieved resynthesis. In much the same way as a pan pot places a source at some position between left and right, the ADRess algorithm will extract a source from some position between left and right.

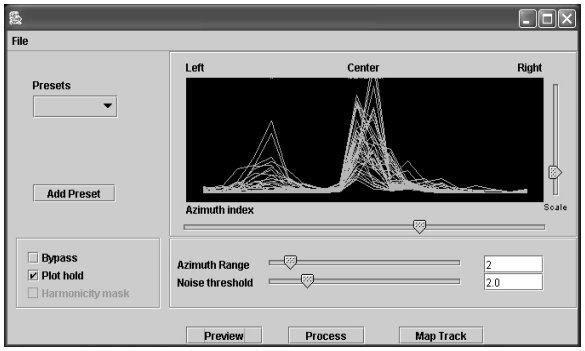

Figure 4: The Interface

The graph in the interface is updated periodically and is obtained using equation 10.

$$E = \left[\sum_{k=1}^{N} A_{ZL(k,i)}\right] \| \left[\sum_{k=1}^{N} A_{ZR(k,i)}\right] F \tag{10}$$

for $0 \le i \le \beta$, where β is the azimuth resolution, F signifies the reverse ordering of the matrix and $\|$ signifies matrix concatenation.

5. RESULTS

We have applied the ADRess algorithm to a number of commercial recordings. The degree of separation achieved depends on the amount of sources, the source proximity and the source level. If sources are proximate, it is likely that multiple sources may get extracted. If there is a large number of sources, partials may migrate towards the source of greatest intensity. If the source level is too low, the resynthesis may have a poor signal to noise ratio. In general though, some degree of separation is possible. The example we have chosen here is an excerpt from a piece of popular jazz music containing saxophone, double bass, drums and piano, all

of which are playing simultaneously. This example and the resultant separations can be downloaded at:

`www.dmc.dit.ie/2002/research_ditme/dnbarry`

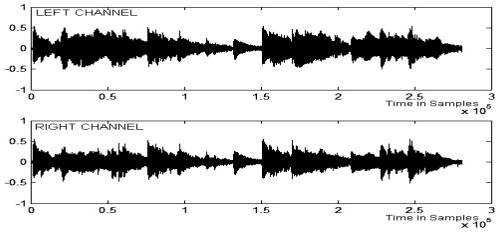

Figure 5: Original stereo recording

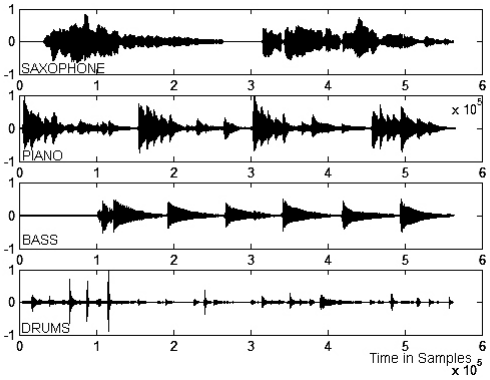

Figure 6: The resultant separations

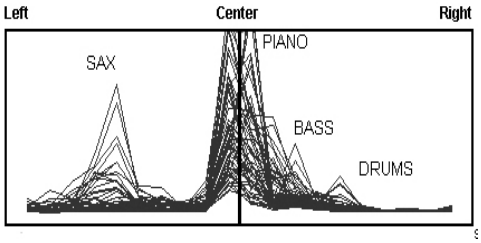

Figure 7: The stereo placement of each instrument

6. CONCLUSIONS

We have presented an algorithm which is capable of performing sound source separation by decomposing stereo recordings into frequency-azimuth subspaces. These subspaces can then be resynthesised individually, resulting in source separation. The only constraints are that the recording is made in the fashion described in section 2, and that the sources do not move position within the stereo field. We feel that ADRess is applicable to a large percentage of commercial recordings. Even in cases where the algorithm cannot achieve complete separation, its ability to reduce the complexity of a mixture could also be deemed useful as a front end for other methods of sound source separation.

7. ACKNOWLEDGEMENTS

Many thanks to Derry Fitzgerald for knowledge imparted and also to Frank Duignan for work on the the real-time java implementation of ADRess.

8. REFERENCES

[1] A. Hyvarinen, J. Karhunen and E. Oja, "*Independent Component Analysis*", Wiley & Sons, 2001.

[2] M.A. Casey, "Separation of Mixed Audio Sources by Independent Subspace Analysis," *Proc. of the int. Computer Music Conference*, Berlin, August 2000.

[3] D.F. Rosenthal, H. G. Okuno, *Computational Auditory Scene Analysis,* LEA Publishers, Mahwah NJ, 1998.

[4] A. Jourjine, S. Rickard, O. Yilmaz, "Blind Separation of Disjoint Orthoganal Signals: Demixing N Sources from 2 Mixtures," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, June2000

[5] S. Rickard, R. Balan, J.Rosca. "Real-Time Time-Frequency based Blind Source Separation," *Proc. of ICA 2001 Conference*, San Diego, CA, December 9-13, 2001.

[6] A.S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[7] J.M. Eargle, "Stereo/Mono Disc Compatibility: A Survey of the Problems," *Journ. of AES*, vol. 17, no.3, pp. 276-281, June1969.

[8] L. Rayleigh, "On Our Perception of Sound Direction," *Phil. Mag*., vol 13, pp. 214-232, 1907.

[9] C. Avendano, J.M. Jot, "Frequency-Domain Techniques for Stereo to Multichannel Upmix," *In Proc. AES 22nd International Conference on*

Virtual, Synthetic and Entertainment Audio, pp. 121-130, Espoo, Finland 2002.

[10] C. Avendano, " Frequency Domain Source Identification and Manipulation In Stereo Mixes for Enhancement, Suppression and Re-Panning Applications," *In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp. 55-58 New Paltz, NY, October 19-22 2003

[11] C. Avendano, J.M. Jot, "A Frequency-Domain Approach to Multichannel Upmix", J. Audio Eng. Soc., Vol. 52, No. 7/8, 2004 July/August