Audio Engineering Society

Convention Paper

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain

Comparison of Signal Reconstruction Methods for the Azimuth Discrimination and Resynthesis Algorithm

Dan BarryP[1], Bob Lawlor[2], and Eugene Coyle[1]

[1] Audio Research Group, Digital Media Centre, Dublin Institute of Technology, Aungier St. Dublin, Ireland
barrydn@eircom.net & eugene.coyle@dit.ie

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

ABSTRACT

The Azimuth Discrimination and Resynthesis algorithm, (ADRess), has been shown to produce high quality sound source separation results for intensity panned stereo recordings. There are however, artifacts such as phasiness which become apparent in the separated signals under certain conditions. This is largely due to the fact that only the magnitude spectra for the separated sources are estimated. Each source is then resynthesised using the phase information obtained from the original mixture. This paper describes the nature and origin of the associated artifacts and proposes alternative techniques for resynthesising the separated signals. A comparison of each technique is then presented

1. INTRODUCTION

The ADRess algorithm [1, 2] 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 estimate the spectra of the separated sources. Although the magnitude spectra are good approximations of the original source spectra, the algorithm makes no attempt at finding a set of phase approximations for source resynthesis. Instead, the phase information taken from the original mixture is used for all sources. This is shown to be acceptable in the majority of cases but artifacts such as phasiness can exist. This is particularly noticeable in percussive or transient audio. Other artifacts can arise when two sources overlapping in the time-frequency domain are

positioned in close proximity to each other in stereo space. These artifacts are the result of what is identified as 'frequency-azimuth smearing' in [2]. Effectively, low energy sources can be significantly degraded by high energy sources in the stereo mixture. For example, a sustained note within in one separation may contain amplitude modulation or even complete dropouts due to the onset of a drum which has been panned to a similar position.

The signal reconstruction in the original ADRess algorithm is achieved by inverting the short-time Fourier Transform (STFT) of the separated source spectra with the original mixture phases. In this paper we explore the use of alternate signal reconstruction methods. Since there is no method for determining the original phase contributions of each source in a mixture, we must rely solely on the magnitude spectra of the separated sources. For this reason, the "magnitude-only" reconstruction technique in [3] is proposed. A Sinusoidal Model [4] resynthesis is also presented here as an alternative reconstruction method. The separated spectra produced by ADRess are simply estimates of the actual source spectra and as such may be distorted, i.e. the lobes associated with peaks in the frequency domain can become smeared which would lead to artifacts on resynthesis. A sinusoidal model reconstruction may provide better results on the basis that only the peaks in the frequency domain are extracted for resynthesis.

2. BACKGROUND

The ADRess algorithm achieves source separation by taking advantage of destructive phase cancellation in the frequency domain. One channel is iteratively gain scaled and subtracted from the other in the complex frequency domain after which the modulus is taken. The resulting array is of dimension N x β, where N is the number of frequency points and β, the azimuth resolution, is the number of equally spaced gain scalars between 0 and 1. The operation reveals local minima, due to phase cancellation, across the azimuth plane for each frequency component. Components belonging to a single source are seen to have their minima in a localised region about some gain scalar which ultimately refers to the pan position of the source in stereo space.

The process can be described as follows; firstly we take the fast Fourier transform (FFT) of a windowed (typically raised cosine) short time segment of length N of each channel,

$$Lf_{(k)} = \sum_{n=0}^{N-1} L_{(n)} W_n^{kn} \tag{1}$$

where $W_n = e^{-j2\pi/N}$ and similarly for the right channel yielding $Lf(k)$ and $Rf(k)$ which represent short time complex frequency representations of the left and right signal. The iterative gain scaling process results in what is termed a 'frequency-azimuth plane' and is constructed using equation 2,

$$AzL_{(k,\,i)} = \left| Rf_{(k)} - g_{(i)}.Lf_{(k)} \right| \tag{2}$$

where $1 \leq k \leq$ N and where $g(i)=i/\beta$, for all i where, $0 \leq i \leq \beta$, and where i and β are integer values. β refers to the number of gain scalars to be used and ultimately give rise to the resolution achieved in the azimuth plane. For example, β=10, will result in 10 discrete azimuth positions for each channel, i.e. 20 positions from left to right. Equation 2 represents the left half of the azimuth plane, $AzL(k,i)$; the right half is created by changing the positions of the left and right variables above. Figure 1 shows the result of the above function for 1 frequency component, k=110,

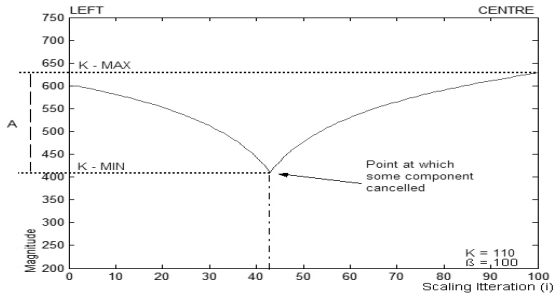

Figure 1: Local minima in bin 110 due to cancellation.

In figure 2, it can be seen that the minima for multiple components from two sources align along the relevant source positions. These local minima represent the points at which frequency components experience a drop in energy due to destructive phase cancellation. This energy drop is directly proportional to the amount of energy which the cancelled source had contributed to the overall mixture and so to invert these minima around a single azimuth point should yield short-time magnitude spectra of the individual sources. To do this inversion we simply subtract the minimum from the maximum of the function as shown in figure 1 and described by equation 2.

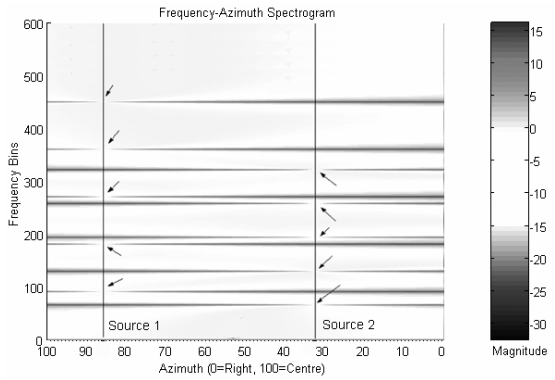

Figure 2: Local minima for 2 complex sources.

To invert the minima we use equation 3.

$$AzR_{(k,i)} = \begin{cases} AzR_{(k)\max} - AzR_{(k)\min}, & \text{if } AzR(k,i) = AzR(k)_{\min} \\ 0, & otherwise \end{cases} \quad (3)$$

The effect of this operation is to turn the minima or nulls into peaks. Equation 3 must be performed for both left and right frequency azimuth planes. At this point we have separated out all frequency components according to the azimuth positions at which they cancelled. It is the case that frequency components and their relative magnitudes relating to a single source will be grouped around a single azimuth position which corresponds to the pan position of the source. In order to resynthesise a source, we simply extract the portion of the frequency azimuth plane around an azimuth position using equation 4,

$$YR_{(k)} = \sum_{i=d-H/2}^{i=d+H/2} AzR(k,i) \qquad \begin{matrix} 1 \le k \le N \\ 0 \le d \le \beta \end{matrix} \quad (4)$$

where *d* is the azimuth index, i.e. the azimuth position of the source for separation and *H* is the azimuth subspace width which is simply a neighborhood around the azimuth index. *YR(k)* is now an N x 1 array containing the short-time magnitude spectrum of a single source or azimuth subspace. Typically at this point, we use an IFFT with the original mixture phases and a standard overlap add technique to resynthesise the signal. One problem is that the estimated spectra no longer have the windowed characteristics of the signal due to the ADRess process. For this reason a synthesis window must also be applied to avoid discontinuities in the resynthesised signal. Furthermore, the overlap is set at 3/4 the frame size (75%) to avoid modulation in the resynthesis since we have effectively windowed the data twice. This reconstruction method gives satisfactory results even though no phase estimates are provided for the separated sources. In the next section, we attempt a reconstruction with only the magnitude spectra which ADRess produces.

3. 'MAGNITUDE ONLY' RECONSTRUCTION

In [3], Griffin et al propose an iterative technique which allows a signal to be reconstructed, given only the modified short-time Fourier transform magnitudes (MSTFTM) and a set of initial, or even random phases. The approach is based on the fact that not all STFTs are 'valid' in the sense that there may not exist a sequence of time values which would yield a given STFT. This is the case with many frequency domain techniques for sound source separation, in that, typically only the magnitude spectra of the sources are estimated. These estimated spectra do not correspond to any 'real' signal. The algorithm in [3] attempts to find a real signal whose STFT is closest in a least squared error sense to the MSTFTM which is provided. Using a standard windowed overlap add procedure, the algorithm iterates between the time and frequency domain; during each iteration the phases are altered due to the influence of 2 consecutive frames overlapping, however, the re-synthesis for any given iteration always uses the original MSTFTM and the updated phases. It is shown by the distance measure described by equation 5, that the squared error between the STFT of the real signal and the MSTFTM is reduced in each iteration. Through this process a set of phase approximations can be arrived at. As the iterations increase, the phase estimates become more accurate until a critical point is reached, after which no significant improvement is achieved.

$$D_i[x^i(n), y_w(mS,\omega)] = \sum_{m=-\infty}^{\infty} \sum_{\omega=-\infty}^{\infty} [\left|x_w^i(mS,\omega)\right| - \left|y_w(mS,\omega)\right|]^2 \quad (5)$$

D_i represents the distance between the STFTM of the resynthesised signal after the i^{th} iteration, $|x_w^i(mS,\omega)|$, and the given MSTFTM, $|y_w(mS,\omega)|$, where *m* is a frame index and *S* is the hopsize.

In equation 5, $x^i(n)$, is notated as such to emphasize the fact that $x_w^i(mS,\omega)$ is a valid STFT, whereas

$y_w(mS,\omega)$ may not be. For the i^{th} iteration then, the resynthesised signal is given by equation 6.

$$x^i(n) = \sum_{\omega=-\infty}^{\infty} \left| y_w(mS,\omega) \right| \cdot e^{j\angle x_w^{i-1}(mS,\omega)}, i > 1 \qquad (6)$$

For the first iteration, i=1, a set of random phases are chosen. The purpose of using this algorithm as a resynthesis method for ADRess was to determine weather a better set of phase approximations could be arrived at than simply using the original mixture phases. The distance measure D_i, given by equation 5, was used to ascertain which set of phase estimates give the best resynthesis in a least squared error sense. Furthermore, the original mixture phases were used as the initial phase estimates for a magnitude only reconstruction to see if the algorithm would converge to even better phase estimates with fewer iterations. Figure 3 shows that the distance is reduced for each iteration where the initial phase estimates are random, but the error is never less than that of simply using the original phases, even after 100 iterations. Informal listening tests suggest that there is no perceivable advantage to using a magnitude only reconstruction and that the original mixture phases provide better results without any iteration than a magnitude only reconstruction with several iterations.

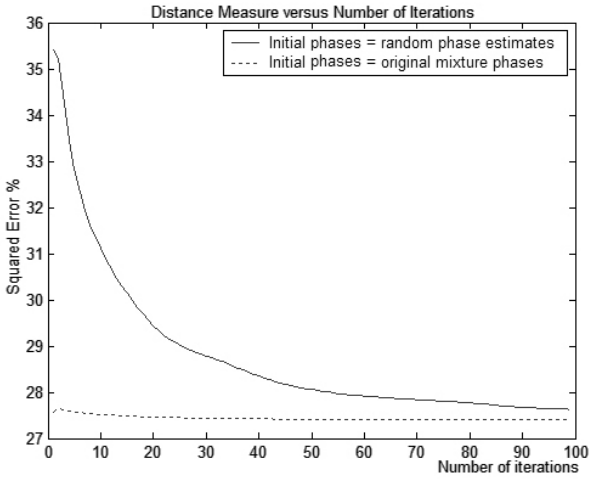

Figure 3: The error reduction as a result of several iterations. Note that the iterative phase estimates never improve on the original mixture phase estimates.

An improved version of the above technique was employed by Slaney for correlogram inversion [6]. The principal difference here is that a synchronized over-lap-add procedure [7] is used to obtain the optimal frame over-lap position to ensure horizontal phase coherence. Ultimately this procedure causes the algorithm to converge with fewer iterations but no perceptual improvement is achieved.

4. SINUSOIDAL MODEL RECONSTRUCTION

Sinusoidal modeling is a well known analysis/synthesis technique for sound modeling and manipulation [4] [5]. The technique is based on the fact that complex musical signals can be represented as a sum of sinusoids with time varying amplitudes, phases and frequencies. These parameters are generally extracted from a time-frequency representation such as the STFT where a sinusoid is represented by a well defined peak with a predictable lobe according to the windowing parameters used in the analysis stage. A peak is usually regarded as any bin with a magnitude greater than that of its two nearest neighbors. The true frequency of the peak can be calculated using either the phase derivative or by using parabolic interpolation. The magnitude is then taken to be the true maximum of the interpolated curve. A peak continuation algorithm tracks peaks from frame to frame to form trajectories. It attempts to find a peak in the next frame with a similar amplitude and frequency to a peak in the previous frame within some threshold of frequency deviation. These frequency, amplitude and phase values are then interpolated to create sinusoidal tracks with time varying amplitudes and frequencies which can easily be synthesized. This is referred to as the deterministic synthesis which corresponds to the steady state harmonic portions of a signal. The deterministic signal can be accurately modeled using only the frequency and amplitude parameters of the interpolated tracks. The 'noise like' or stochastic parts of the signal can be estimated by subtracting the deterministic signal from the original signal; in this case however, the deterministic synthesis must contain the instantaneous phase values obtained in the analysis stage. The residual which is assumed to be stochastic, is then usually modeled as time varying filtered noise. The basic sinusoidal model architecture has been described here but there are many heuristics which control the behavior of the peak continuation algorithm. One such heuristic gives us the ability to discard sinusoidal tracks which are shorter than a specified duration. This is of particular interest to us since the separations achieved with the ADRess algorithm are subject to brief interference from neighboring sources. This sort of interference as well as noise, appears as 'speckling' on

the spectrogram of the separated source The ability to remove trajectories with such short duration should allow a cleaner resynthesis of the deterministic parts of the signal.

Here we use a modified sinusoidal model implemented by Ellis [8] to carry out the resynthesis of the separated source spectra generated by the ADRess algorithm. The sinusoidal modeling technique is quite flexible, but this flexibility comes at a cost; adjusting the algorithm parameters for optimal performance depends largely on the signal characteristics and so configuring the algorithm can be quite tedious. For the example shown in figure 4, the algorithm was configured in such away as to reject as much noise and neighboring source interference as possible. Trajectories with durations less than 6 frames were also discarded. The source in this case was a saxophone which has been separated from a mixture of piano, bass, saxophone and drums.

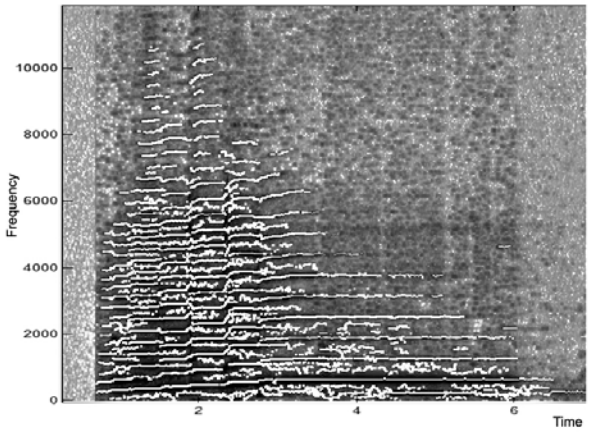

Figure 4: Trajectories (shown in white) formed by the peak continuation algorithm superimposed over the spectrogram returned by the ADRess algorithm.

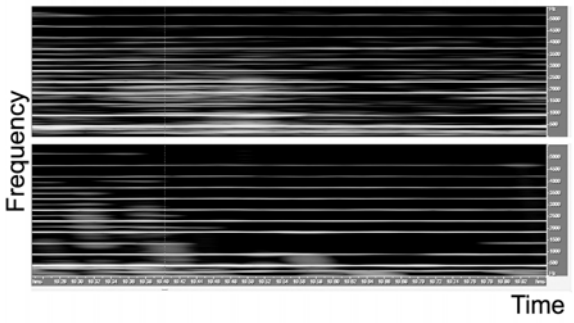

Figure 5: Close up on the spectrogram of a pitched region of the saxophone separation with the standard iSTFT method shown on top and the sinusoidal model on bottom.

The sinusoidal model resynthesis although cleaner in the pitched regions suffers from artifacts when parameters are incorrectly set. The task of determining how much of the residual signal belongs to the signal and how much is unwanted noise can be difficult, making threshold setting very much a trial and error procedure. However, the results are compelling, and the sinusoidal model could be adapted for the purposes of an offline resynthesis.

5. CONCLUSIONS

We have explored the use of two alternative reconstruction techniques for the ADRess algorithm. Firstly the magnitude only reconstruction technique was applied to the separation spectra produced by ADRess in an attempt to arrive at a set of suitable phase estimates. Although the error is reduced significantly after 50 iterations or so using random phase estimates, the error between the initial spectrogram and the final spectrogram is never less than that when the original mixture phases are used. We believe that the reason for this is linked to a condition identified by Rickard et al known as W-disjoint orthogonality [9]; two sources are said to be W-disjoint orthogonal if there is no significant overlap between the sources time-frequency representations. In the case of musical signals there is usually quite significant overlap in frequency and time, this overlap is the cause of what is identified as ‘frequency-azimuth smearing’ in [1]. Effectively when multiple sources contribute to a single frequency component, their phase contributions cause phase cancellation errors in the ADRess algorithm; this in turn causes the frequency dependent nulls to drift away from the apparent azimuth position of a particular source. Sources with the highest intensity will have most influence over the resultant phases when sources are mixed, and as such will be separated better by ADRess. Furthermore, the phases for any time-frequency point of a mixture of sources will be closest the phase of the source with the greatest magnitude at that time-frequency point. This leads us to the assumption that there is a variable W-disjoint orthogonality associated with musical mixtures which is purely dependent on the mixture at any given point in time. So for points in time where the sources do not overlap significantly in the

frequency domain, the original mixture phases are a close approximation to the source phases.

A sinusoidal model was also applied as a resynthesis technique for the separated source spectra. The technique does offer some advantages for the synthesis of deterministic signals in that some noise and source interference can be rejected resulting in cleaner resynthesis of pitched regions of the signal. The primary disadvantage is that the technique requires that the operational parameters of the algorithm need to be adjusted accordingly depending on the signal.

The ADRess algorithm has been implemented to run in real-time and so computational efficiency is particularly important. Although the reconstruction methods explored here are useful, the method of using the original mixture phases with a standard inverse STFT is still the preferred option as it gives the best trade-off between quality and efficiency.

6. ACKNOWLEDGEMENTS

Thanks to Derry Fitzgerald, David Dorran and Frank Duignan for code and communication.

7. REFERENCES

[1] Barry, D., Lawlor, R. and Coyle E., "Sound Source Separation: Azimuth Discrimination and Resynthesis", *Proc. 7th International Conference on Digital Audio Effects, DAFX 04*, Naples, Italy, 2004

[2] Barry, D., Lawlor, R. and Coyle E., "Real-time Sound Source Separation using Azimuth Discrimination and Resynthesis", *Proc. 117th Audio Engineering Society Convention*, October 28-31, San Francisco, CA, USA, 2004.

[3] Griffin D. W., Lim J.S., "Signal Estimation from Modified Short-Time Fourier Transform", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-32, no. 2, April 1984

[4] McAulay, R.J. and T.F. Quatieri. "Speech Analysis/Synthesis based on a Sinusoidal Representation", *IEEE Transactions on Acoustics, Speech and Signal Processing* 34(4):744--754. 1986.

[5] Serra, X, "Musical Sound Modeling with Sinusoids plus Noise", published in Roads, C., Pope, S., Picialli, A., De Poli, G., editors. "Musical Signal Processing". Swets & Zeitlinger Publishers.

[6] Slaney, M., Naar, D., and Lyon, R.F., "Auditory Model Inversion for Sound Separation", *Proc. ICASSP 94 – 1994 International Conference on Acoustics, Speech, and Signal Processing,* Adelaide, Australia, 19-22 April 1994.

[7] S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," *Proc. of the IEEE ICASSP*, 493-496, 1985.

[8] Ellis, D., *Matlab implementation of a sinusoidal model.* http://www.ee.columbia.edu/~dpwe/resources/matlab/sinemodel/

[9] Rickard, S., Yilmaz, O., "On the approximate W-disjoint orthogonality of speech", 529–532, ICASSP 2002.