2008-06-01

Automatic Detection of Optimal Azimuth Widths for Sound Source Separation using Adress

Dan Barry
Technological University Dublin, dan.barry@tudublin.ie

Derry Fitzgerald
Cork Institute of Technology

Matt Cranitch
Cork Institute of Technology, matt.cranitch@cit.ie

See next page for additional authors

Recommended Citation

Barry, D., Fitzgerald, D., Cranitch, M. & Coyle, E. Automatic detection of optimal azimuth widths for sound source separation using Adress. Paper given at the *Irish Signals and Systems Conference, Galway, June 18-19, 2008.* http://www.audioresearchgroup.com/

Automatic detection of optimal azimuth widths for sound source separation using Adress.

Derry FitzGerald[†], Dan Barry[*], Matt Cranitch[†] and Eugene Coyle[*]

[†]*Dept. of Electronic Engineering*
Cork Institute of Technology, Cork

[*]*School of Electrical Engineering Systems*
Dublin Institute of Technology, Dublin

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

Abstract — **The Adress algorithm has proved successful as a means of separating sound sources from stereo mixtures. The algorithm has two main parameters, azimuth position and azimuth width, and these are typically set by the user for each source individually. For the separation of large amounts of audio material, such as in an audio archive, a method of automatically setting these parameters would be of use. This paper proposes a method of automatically obtaining the azimuth widths for the sources in a mixture by balancing the reconstruction error between the original spectrograms and the resynthesised spectrograms with the sparseness of the recovered sources, using an L-curve type approach.**

Keywords — **Sound Source Separation, Sparsity**

I The Adress Algorithm

The Adress algorithm was developed for the purpose of sound source separation from linear stereo mixtures where different sources are positioned at different positions in the stereo field [1]. The signal model underlying Adress can be described as:

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

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

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

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

As the mixture signals are linear mixtures of the original sources, it can be seen that $L - g_j R$ will cause the jth source to be cancelled from the mixture. While this is sufficient to eliminate a given source from a mixture of sources, recovery of the cancelled source requires the use of frequency domain techniques.

Therefore, a Short Time Fourier Transform (STFT) is carried out on each of the mixture signals, using the same parameters for each signal, typically a 4096 point FFT, with a hopsize of 1024 points. A frequency-azimuth plane is then created for each channel, where β is the azimuth resolution, which determines how many equally spaced gain scaling values of g are used to create the plane. β is related to the gain as follows

$$g_i = i * \frac{1}{\beta} \tag{4}$$

where $0 \leq i \leq \beta$ and where i and β are integers. A frequency-azimuth plane for each channel can then be defined as:

$$Az_{R(k,i)} = |Lf_k - g_i Rf_k| \tag{5}$$

$$Az_{L(k,i)} = |Rf_k - g_i Lf_k| \tag{6}$$

where Az_R and Az_L are the right and left channel azimuth frequency planes, and Rf_k and Lf_k are the current right and left frames of the STFT respectively.

At azimuth positions where a source is present, the energy in the frequency bins associated with a given source will be cancelled out and there will be a minimum at that position in the azimuth frequency plane. In order to estimate the magnitude of the energy lost due to cancellation at each frequency bin, the azimuth-frequency plane is then redefined as:

$$Az_{R(k,i)} = \begin{cases} Az_{R(k)_{max}} - Az_{R(k)_{min}} & \text{if} Az_{R(k,i)} = Az_{R(k)_{min}} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$Az_{L(k,i)} = \begin{cases} Az_{L(k)_{max}} - Az_{L(k)_{min}} & \text{if} Az_{L(k,i)} = Az_{L(k)_{min}} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

To separate sources at different spatial positions in the stereo field, a discrimination index d is defined, where $0 \leq d \leq \beta$, with d indicating the position of the source in the azimuth plane. In many cases dealing with musical sources, there will be frequency overlap between different sources which will cause the position of a frequency minimum to drift away from that of the source position. In order to overcome this problem, an azimuth subspace width, H is defined, so that $1 \leq H \leq \beta$. Together with d, this defines what portion of the azimuth-frequency plane to resynthesise.

Given d and H for a given source, the magnitude spectrogram of the current frame of that source can be estimated from:

$$Y_{R(k)} = \sum_{i=d-H/2}^{i=d+H/2} Az_{Rk,i} \quad (9)$$

$$Y_{L(k)} = \sum_{i=d-H/2}^{i=d+H/2} Az_{Lk,i} \quad (10)$$

A time-domain signal can be obtained by applying the phase information from the original mixture signal to the magnitude spectrogram and carrying out an inverse Fourier Transform.

The Adress Algorithm has been shown to give high quality separations for stereo mixtures, and is capable of functioning in real-time [2]. However, to-date, no research has been carried out on techniques for determining the optimal values of d and H for separation of the sources in a mixture. This would be useful in cases where a large volume of audio material has to be separated, such as in a large audio archive. The remainder of this paper focuses on developing a technique for the determining the optimal values of H for the sources in a given mixture.

II Sparsity and the L-curve

In the spectrograms of any mixture of sound sources there will be a greater number of time-frequency bins with significant energy in them than in the individual source spectrograms. Therefore it can be seen that the spectrograms of individual sources will be sparser than the mixture spectrograms. It follows from this that a good solution to the problem of separating sound sources is to find sparse individual source spectrograms, which, when combined, still give a good reconstruction of the original mixture spectrograms. It can be appreciated that there is a trade-off between the sparseness of the sources and the reconstruction of the mixtures, and a technique is needed to measure this trade-off.

A simple approach to this problem is the L-curve method proposed by Hansen [3]. Though the L-curve was developed for use in another context, it has been found by Morup et al. to generalise well for use in measuring the trade-off between sparseness and reconstruction error [4]. In this paper, the L-curve was used in the context of sparse coding. A typical cost function used in sparse coding is

$$D_{spar}(\mathbf{A}, \mathbf{S}) = \frac{1}{2}\|\mathbf{X} - \mathbf{AS}\|^2 + \lambda \sum_{i,} |S_{ij}| \quad (11)$$

where $\mathbf{X}$ is the original data matrix and $\mathbf{A}$ and $\mathbf{S}$ are the matrix factors used to reconstruct $\mathbf{X}$. The cost function attempts to balance the reconstruction of the original data, as measured by the the squared Euclidean distance given in the first term, with the sparseness of the reconstruction, as measured by the second term. The sparseness of the solution is determined by the value of λ.

The L-curve was then obtained by plotting the reconstruction error, against the $l0$-norm of the sparse code matrix for various values of λ. The optimal value of λ was taken to be the value associated with the point of maximum curvature of the L-curve. This method was found to give good results both in image analysis and in the sound source separation of musical instruments.

In the context of the Adress algorithm, once the azimuth positions of the sources have been determined, the sparseness of the solution is determined by the azimuth width H, which controls the number of frequency bins used to reconstruct a given source, with the fewer bins used, the sparser the representation of the recovered source. Therefore, it is proposed to use an L-curve approach to determine the optimal azimuth widths that balance reconstruction of the original mixture spectrograms with the sparseness of the source spectrograms.

Unlike the sparse coding case where there is a single parameter λ to be optimised, here the number of parameters to be optimised is equal to the

number of sources identified in the mixture. This makes plotting an L-curve directly unfeasible, and so we have adopted the approach of using an L-curve distance measure defined as:

$$D_L = \sqrt{r^2 + l0^2} \tag{12}$$

where r is the reconstruction error and $l0$ is the $l0$-norm for a given set of azimuth widths, with one width per source. The optimal set of azimuth widths is then determined as the set of widths associated with the smallest D_L.

III Measures for Data Reconstruction

The L-curve approach requires the use of a measure of the reconstruction error between the original data and the reconstructed data. In the context of sparse coding the squared Euclidean distance was used, which measures the closeness of the reconstruction in a least squares sense:

$$D_{SED}\left(\mathbf{X}, \hat{\mathbf{X}}\right) = \frac{1}{2}\sum_{ij}(x_{ij} - \hat{x}_{ij})^2 \tag{13}$$

However, in the context of sound source separation algorithms, it has been found that other measures often give better results. In particular, the generalised Kullback-Liebler (KL) divergence, which has found widespread use in non-negative matrix factorisation [5], has been found to be a useful measure for sound source separation of musical instruments [6]. In light of this, it was decided to use this distance to see if it gave better performance in this context than the squared Euclidean distance. The generalised KL divergence is given by:

$$D_{KLD}(\mathbf{X}\|\hat{\mathbf{X}}) = \sum_{ij}\left(x_{ij}log\frac{x_{ij}}{\hat{x}_{ij}} - x_{ij} + \hat{x}_{ij}\right) \tag{14}$$

Another measure used in testing the L-curve approach is the Itakura-Saito divergence [7], which was designed as a similarity measure for speech signals, as it was felt that it might have some applicability in a musical context as well. The Itakura-Saito (IS) distance is given by:

$$D_{ISD}(\mathbf{X}\|\hat{\mathbf{X}}) = \sum_{ij}\left(\frac{x_{ij}}{\hat{x}_{ij}} - log\frac{x_{ij}}{\hat{x}_{ij}} - 1\right) \tag{15}$$

IV Source Separation Performance Metrics

In order to quantitatively measure the quality of the separations obtained, a set of separation performance metrics must be used. A commonly used set of metrics are those defined by Vincent et al [8]. Here the recovered time domain signal is decomposed into the sum of three terms, with reference to the original unmixed source signal:

$$s_{rec} = s_{tar} + e_{int} + e_{art} \tag{16}$$

where s_{rec} is the recovered source signal, s_{tar} is the portion of the recovered signal that relates to the original or target source, e_{int} is the portion that relates to interference from other sources, and e_{art} is the portion that relates to artifacts generated by the separation technique and/or the resynthesis method. Based on this decomposition, source separation metrics were then defined.

The first of these, Signal to Distortion ratio (SDR), provides a measure of the overall quality of the sound source separation:

$$SDR = 10log_{10}\frac{\|s_{tar}\|^2}{\|e_{int} + e_{art}\|^2} \tag{17}$$

The Signal to Interference ratio (SIR) provides a measure of the presence of other sources in the separated source:

$$SIR = 10log_{10}\frac{\|s_{tar}\|^2}{\|e_{int}\|^2} \tag{18}$$

Finally, the Signal to Artifacts ratio (SAR) provides a measure of the artifacts present in the signal due to separation and/or resynthesis:

$$SAR = 10log_{10}\frac{\|s_{tar} + e_{int}\|^2}{\|e_{art}\|^2} \tag{19}$$

These metrics are invariant to scaling factors and were calculated using the BSS_EVAL toolbox available at [9].

V Evaluation Methodology

In order to evaluate the performance of the L-curve approach in determining the optimal azimuth widths for source separation, a set of 15 test signals were created using a large library of orchestral samples [10]. The 15 test signals were of 4 seconds duration and contained mixtures of melodies played by three different instruments or sources. Samples from a total of 15 different orchestral instruments were used. Fixed azimuth positions were used for the sources in each example, with the first source at -0.6, the second at 0 and the third at 0.6, where an azimuth postion of -1 corresponds to fully left, zero to the centre and 1 to fully right. These azimuth positions were provided to the Adress algorithm.

For each test signal, the azimuth width for each of the sources was varied from 0.05 to 0.5 in steps of 0.05, resulting in 10 azimuth widths for each source. The source spectrograms for each azimuth width of each source were obtained, and for each of the 1000 possible combination, the $l0$-norm, D_{SED}, D_{KLD} and D_{ISD} were obtained. From these D_L, the L-curve distance measure between the $l0$-norm and the data reconstruction measures was calculated.

For each azimuth width of each source, the resultant source spectrogram was inverted back to the time domain by applying the original mixture phase information to the spectrogram. The phase of the left channel used for left dominant sources, and the phase of the right channel used for right dominant sources, with the left channel phase information used for sources positioned in the centre. SDR, SIR and SAR were then calculated from the recovered time-domain signals for each width of each source. For each of the 1000 possible combinations of the 3 sources, an overall SDR, SIR and SAR were calculated by taking the mean of the metrics for the individual sources. This was done in order to provide an overall measure of separation for each of the azimuth width sets. The results obtained are discussed in the following section.

VI Results

On examination of the results obtained, it was found that a high degree of correlation was observed across all test signals between the L-curve distance measures for the generalised Kullback-Leibler distance and both SDR and SAR, and similarly for the Itakura-Saito distance. No correlation was observed with the squared Euclidean distance, and no correlation was observed with SIR and any of the distance measures. Figure 1 shows the results obtained for one of the test signals. The 1000 datapoints corresponding to all possible azimuth width combinations were sorted by their SDR, and the plot shows SDR, $-D_L$ for the KL distance, $-D_L$ for the IS distance, and $-D_L$ for the squared Euclidean distance. It can be seen that the trends for the KL and IS distances closely follow those of the SDR, while no relationship is visible for the squared Euclidean distance.

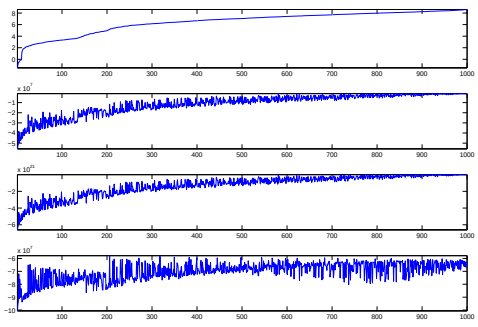

Fig. 1: Results obtained from one of the test signals for SDR, $-D_L$ for the KL distance, $-D_L$ for the IS distance, and $-D_L$ for the squared Euclidean distance respectively

Similarly, Figure 2 shows the results obtained for SIR and the distance measures, sorted by SIR. No correlation can be seen between SIR and any of the measures. The results obtained for SAR were very similar to those for SDR and so are not shown.

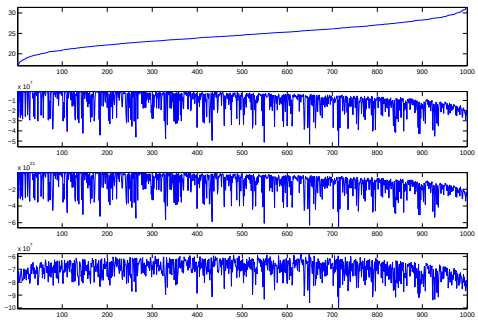

Fig. 2: Results obtained from one of the test signals for SiR, $-D_L$ for the KL distance, $-D_L$ for the IS distance, and $-D_L$ for the squared Euclidean distance respectively

Figure 3 shows the correlation coefficients obtained between SDR and the 3 distance measures for each of the 15 test signals. Results for the KL distance are shown as a solid line with a cross indicating the data points, the IS distance is shown as a dash-dotted line with a square marking the data points, and the squared Euclidean distance is shown as a dashed line with circles marking the points. There is a large negative correlation between SDR and both the KL and IS distances, while no significant correlation is observed for the squared Euclidean distance. It can also be seen that the results for both KL and IS distances are very similar, though the KL distance does on average outperform the IS distance. Further, the results for SAR were very similar to those for SDR and so no figure is shown for SAR.

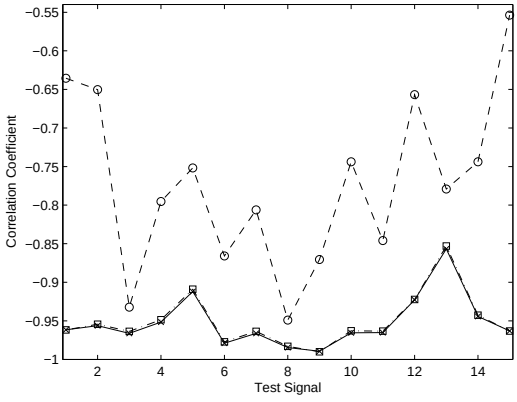

Fig. 3: Correlation Coefficients obtained for SDR with D_L for the KL distance (solid line, cross for data points), D_L for the IS distance (dash-dotted line, square for data points), and D_L for the squared Euclidean distance (dashed line with circle for data points) respectively

Figure 4 shows the correlation coefficients for SIR and the distance measures. No significant cor-

relation is shown between SIR and each of the measures.

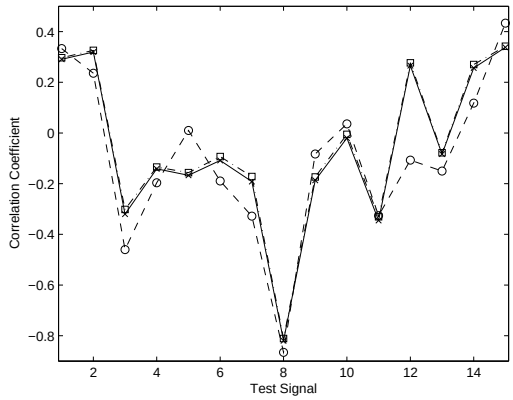

Fig. 4: Correlation Coefficients obtained for SiR with D_L for the KL distance (solid line, cross for data points), D_L for the IS distance (dash-dotted line, square for data points), and D_L for the squared Euclidean distance (dashed line with circle for data points) respectively

It can be seen from the above that D_L for both KL and IS distances can be used to provide a good estimate of which sets of azimuth widths should be used to obtain high SDR and SAR values, with lower D_L values providing higher SDR and SAR values. Figure 5 shows the maximum SDR scores obtained for each of the data signals, as well as the maximum SDR scores obtained by using the set of azimuth widths associated with the minima of each of the reconstruction measures. It can be seen that the KL distance provides SDR scores which are closer to the actual maximum scores than the IS distance, both of which considerably outperform the squared Euclidean distance, and that in all cases the difference between the actual minimum score and that of the L-curve distance method is less than 1 dB for the KL distance, with an average difference of 0.35 dB. The average difference is 0.83 dB and 3.1 dB for the IS distance and the squared Euclidean distance respectively. Similar results are obtained for SAR.

Figure 6 shows the maximum SIR scores obtained for each of the data signals, as well as the maximum SIR scores obtained by using the set of azimuth widths associated with the minima of each of the reconstruction measures. It can be seen that in general there is a large difference between the maximum SIR and those returned using the L-curve method. However, it can also be seen that the lowest SIR scores returned using the L-curve method with the KL distance are all higher than 22 dB. This is still a very high level of rejection of the other sources, and this is evident on listening to the resynthesised source signals, with little or no evidence of other sources in the recovered signals.

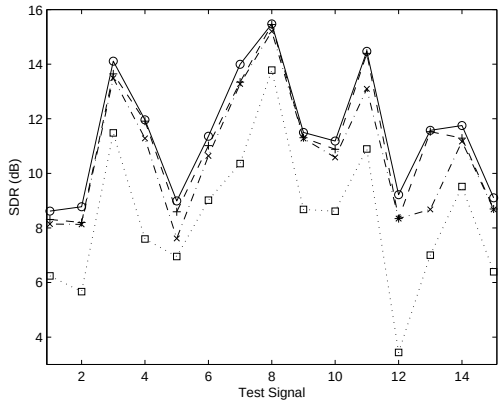

Fig. 5: SDR values for each test signal, maximum achieved (solid line, circle for data points), SDR for minimum KL distance (dashed line, + for data points), SDR for minimum IS distance (dash-dotted line, x for data points), and SDR for the squared Euclidean distance (dotted line, square for data points) respectively

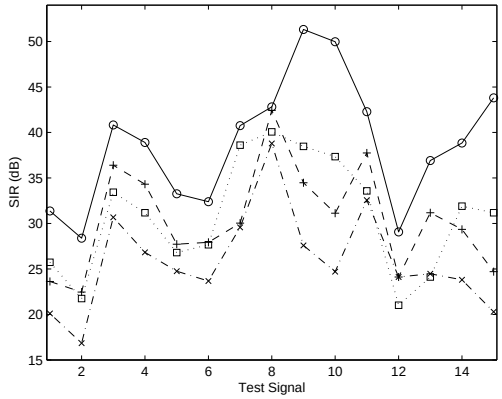

Fig. 6: SIR values for each test signal, maximum achieved (solid line, circle for data points), SIR for minimum KL distance (dashed line, + for data points), SIR for minimum IS distance (dash-dotted line, x for data points), and SIR for the squared Euclidean distance (dotted line, square for data points) respectively

Further, the results for SDR and SAR are very similar, which suggests that the limiting factor in the resynthesis quality of the separated sources is not interference from other sources, but artifacts from the separation and resynthesis, or in terms of Eqn. 16 that $e_{art} \gg e_{int}$ and so SDR becomes approximately equal to SAR. Therefore, once SIR is high, as it is in the sources recovered using the L-curve with KL method, the more important consideration is high SDR. In this light, the L-curve distance method, when used in conjunction with the KL distance, can be seen as a good approximation to the optimal overall separation of the sources in the mixture. This is evident in listening to the recovered source signals.

However, there is a downside to the method,

in that it currently requires an exhaustive search of all combinations of azimuth widths across all sources. While this still only takes a couple of minutes, it considerably slows down recovery of the sources.

VII Conclusions

The Adress sound source separation algorithm was described, and the fact that the azimuth position and width had to be manually fixed highlighted. It was then noted that there was a trade-off between the sparseness of the sources, as controlled by the azimuth width and accurate reconstruction of the original mixture spectrograms, and that the sparsest set of recovered sources which still gave good reconstruction were likely to be a good approximation to the actual sources.

Taking inspiration from the L-curve method used in sparse coding to control the trade-off between sparseness and accurate reconstruction, the method was adapted to deal with determining the optimal set of azimuth widths, by obtaining a distance measure combining the $l0$-norm of the recovered spectrograms with the reconstruction error. A number of different reconstruction measures were used and it was found that the generalised Kullback Leibler distance gave the best approximation to the optimal separation of the sources as measured by SDR, with results close to the optimal obtained consistently.

However, a drawback of the method is that, at present, it requires an exhaustive search of the possible combinations of azimuth widths across all the sources. Future work will concentrate on developing faster methods of performing this search.

VIII Acknowledgements

This research was supported by Enterprise Ireland under the Commercialisation Fund.

References

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

[2] Barry, Dan; Coyle, Eugene; Lawlor, Bob; "Real-time Sound Source Separation using Azimuth Discrimination and Resynthesis", Proc. 117th Audio Engineering Society Convention, Moscone Centre, San Francisco, CA, USA, 2004

[3] P. C. Hansen. "Analysis of discrete ill-posed problems by means of the l-curve". SIAM Review, 34(4):561-580, 1992.

[4] M. Morup, M. N. Schmidt, L. K. Hansen, "Shift Invariant Sparse Coding of Image and Music Data", submitted, JMLR, 2007

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

[6] FitzGerald, D., Cranitch, M., and Coyle, E.,"Shifted 2D Non-negative Tensor Factorisation", Proceedings of the Irish Signals and Systems Conference, Dublin, June 2006.

[7] Itakura, F. and Saito, S."An analysis-synthesis telephony based on maximum likelihood method". In 6th Int. Conf. Acoustics, pages 1720, 1968.

[8] E. Vincent, R. Gribonval and C. Fvotte. "Performance measurement in Blind Audio Source Separation", IEEE Trans. Audio, Speech and Audio Processing, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.

[9] BSS_Eval toolbox available at http://bass-db.gforge.inria.fr/bss_eval/

[10] Peter Siedlaczek, Advanced Orchestra Library Set, 1997.