Audio Engineering Society
Convention Paper
Presented at the 119th Convention
2005 October 7–10 New York, New York USA

Single Channel Source Separation using Short-time Independent Component Analysis

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

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

[2] Dept. Electrical Engineering, Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland
derry.fitzgerald@cit.ie

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

ABSTRACT

In this paper we develop a method for the sound source separation of single channel mixtures using Independent Component Analysis within a time-frequency representation of the audio signal. We apply standard Independent Component Analysis techniques to contiguous magnitude frames of the short-time Fourier transform of the mixture. Provided that the amplitude envelopes of each source are sufficiently different, it can be seen that it is possible to recover the independent short-time power spectra of each source. A simple scoring scheme based on auditory scene analysis cues is then used to overcome the source ordering problem ultimately allowing each of the independent spectra to be assigned to the correct source. A final stage of adaptive filtering is then applied which forces each of the spectra to become more independent. Each of the sources is then resynthesised using the standard inverse short-time Fourier transform with an overlap add scheme.

1. BACKGROUND

Sound source separation has been the topic of extensive research in recent years. The problem has seen many different formulations which have been based on many different mixing models. Some success has been had for the degenerate case, i.e. more sources than mixtures, where there are at least 2 mixture signals. In [1], a technique for recovering N sources from two convolute speech mixtures was presented. The technique clusters time-frequency components based on the intensity ratios and phase delays between each sensor capturing the mixture. In [2, 3] a similar technique was proposed for separating N sources from intensity panned stereo recordings. It is a localisation technique based on clustering components belonging to phase coherent sources emanating from the same position in the horizontal plane. Standard ICA [4] techniques have proved ideal for the case where a number of linear mixtures equal to the number of sources exists.

Unfortunately, the most popular commercial music formats are still of the stereo variety and so standard ICA is rarely applicable. By far the most difficult problem is that of separating N sources from a single mixture of the sources. Some limited success has been achieved to this end using computational auditory scene analysis (CASA) techniques [5] which usually require prior knowledge of some description. Independent Subspace Analysis (ISA) [6] has been applied quite successfully to the separation of drum sounds from single channel mixtures [7]. This technique involves carrying out Principal Component Analysis (PCA) on a magnitude spectrogram which results in a set of N time basis functions (amplitude envelopes) and N frequency basis functions (spectra). They are ordered by variance. The basis functions are de-correlated from each other but not mutually independent. At this stage ICA can be applied to the time basis functions resulting in independent amplitude envelopes which usually correspond to each of the drums in the mixture. The process works well on drums because they tend to account for most of the variance in musical signals. However, because of the way in which the model represents the data, it is limited to pitch stationary sounds such as drums. In this paper we present a method for performing single channel source separation using a combination of ICA and CASA.

2. SYSTEM OVERVIEW

ICA is a statistical method used for blind source separation. It is usually applied to a matrix which contains a set of linear time mixtures of some latent sources. Furthermore, it requires at least as many mixture signals as there are sources present in the mixture. For this reason it is not normally associated with single channel separation since to separate *N* sources you would require *N* observation mixtures. For a detailed description of ICA refer to [4]. Here, we present a formulation which allows ICA to be performed on time-frequency representations of single channel mixtures. The process starts by taking the magnitude STFT of a single channel mixture. Each musical source will usually have significantly different amplitude envelopes which are ultimately a function of the timbre, mode of excitation and articulation possibilities of specific instruments. Therefore, if two instruments are playing simultaneously, the mixture at time (*t*) will be similar to that of the mixture at time (*t+a)* but the mixing coefficients of each instrument at time (*t*) and (*t+a)* will be different because of their time evolving amplitude envelopes. Although the time representations of the signal at *t* and (*t+a*) are not candidates for ICA, the short-term power spectra are.

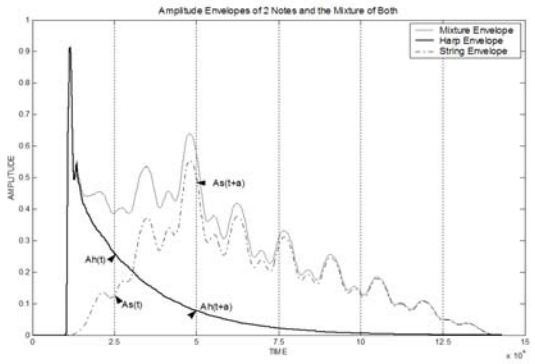

Figure 1: The graph shows the amplitude envelopes of two different notes of equal length along with the resultant mixture envelope of the notes. The dark line plot is the envelope of a harp pluck while the dashed line is that of a bowed string.

In the diagram above, the pitch of each note remains constant but the amplitude coefficients or envelopes of each note change with respect to time. *As(t)* and *Ah(t)* in figure1 refer to the amplitudes of each source, harp and string respectively, at time *(t).* It is effectively these coefficients which are the latent mixing coefficients we seek to discover. These amplitude changes are of course reflected in the time-frequency domain. So a spectral frame at time (*t*) and (*t+a*) can be considered as alternative linear mixtures of the same latent data, i.e. the latent short-time spectra of each source.

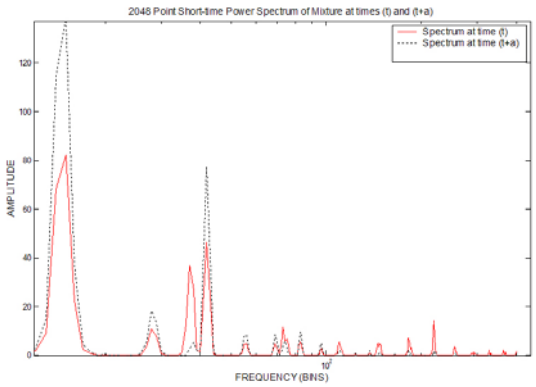

Figure 2: Shows the mixture spectra at times (*t*) and (*t+a*).

In order to separate the sources then, contiguous frequency frames separated by some distance, *a*, are passed to an ICA algorithm which returns independent short-time spectra which correspond to each of the

independent sources. For this reason we have termed the technique, 'Short-time Independent Component analysis' (STICA). After the ICA stage, some processing takes place which ensures that the independent magnitude frames are scaled and ordered correctly. Each independent magnitude spectrum is then synthesized using the IFFT with the original mixture phase information resulting in the separation of two sources from one mixture.

3. METHOD

3.1. ICA Front-end

We begin by taking the magnitude STFT of the mixture, eq. 1,

$$X(k,m) = abs\left[\sum_{n=0}^{N-1} w(n)\,x(n+mH)\,e^{-j2\pi nk/N}\right] \quad (1)$$

where *X(k,m)* is the absolute value of the complex STFT and where *m* is the time frame index, *k* is the frequency bin index, *H* is the hopsize between frames and *N* is the FFT window size. w(n) is a suitable window of length *N* also. Next we perform ICA on 2 contiguous frames of the magnitude spectrum *X(k,m).* The idea is that, over a short time period (1-4 frames), the local frequency content of the signal will be similar but not remain stationary. Therefore, the time varying amplitude envelopes of each source will dictate their relative amplitudes in the mixture at any given time. These unknown amplitude envelopes values of each source at any given time frame are considered to be the mixing matrix denoted by **A** in equation 1. So in order to retrieve 2 independent short-time source spectra we need to supply the ICA algorithm with 2 short-time mixture spectra. Furthermore, to ensure convergence, the mixing coefficients of each source at each time frame must not be the same. So for example, if the amplitude envelopes of the sources change rapidly, consecutive frames may be used but if the envelopes evolve slowly it may be necessary to put some distance between the frames, this distance is denoted by *a* in equation 3 below. It should also be noticed then, that if each of the sources have identical amplitude envelopes, no separation can be achieved. The formal ICA representation is shown in equation 2, where **x** is a set of observation mixtures, **A** is an unknown mixing matrix and **s,** also unknown**,** is a matrix with the same dimension of **x** and contains the independent short-time source spectra.

$$\mathbf{x} = \mathbf{A}\mathbf{s} \quad (2)$$

where,

$$\mathbf{x} = (X(k,m), X(k,m+\alpha)) \quad (3)$$

and where,

$$\mathbf{s} = (S_i(k,m), S_j(k,m)) \;\; i \neq j \quad (4)$$

Referring specifically to a 2 source problem as in equation 3 and 4, both **x** and **s** are of dimension 2 x *k* and **A** is of dimension 2 x 2. *a* is the distance between the frames to be passed to the ICA algorithm. So S_i and S_j then, are the unknown independent short-time source spectra. There are two significant issues at this point. The first issue is that ICA causes the independent outputs to have an arbitrary scaling and the second problem is that ICA returns the independent sources in a random order. This effectively means that source *i* in frame 1 may be source *j* in frame 2. So in order to resynthesise 2 coherent independent sources in time, some method of grouping each frame with the correct source is required. This will be dealt with in section 3.3.

3.2. Re-scaling Frames

As stated, the outputs of the ICA algorithm are always arbitrarily scaled. This would result in random gain changes at the output. The problem is easily resolved since the sum of the independent components should be approximately equal to a scaled version of the original mixture frame. So in order to calculate each source's relative magnitude we use equations 5 and 6.

$$S_1'(k,m) = X(k,m)\frac{|S_1(k,m)|}{|S_1(k,m)| + |S_2(k,m)|} \quad (5)$$

$$S_2'(k,m) = X(k,m)\frac{|S_2(k,m)|}{|S_1(k,m)| + |S_2(k,m)|} \quad (6)$$

The process described above ensures that the scaling between frames is relative and not arbitrary, thus avoiding random gain changes in the output signals.

3.3. Source Ordering

The second problem is that ICA returns the independent sources in a random order and since we are carrying this out on a short-time basis, there is an issue associated with identifying each of the sources so that each of the independent frames are assigned to the correct source for resynthesis. The complexity of this problem rises

significantly the more sources there are present. We have limited this research to only two sources present in a single mixture. A simple scheme to order the sources using 3 measures is proposed. For our examples we use three measures:

1. Normalised Spectral Centroid,
2. Peak Location,
3. Proximity to Previous Peak Locations.

The measures are compared and the results summed to form likelihood scores which determine which sources the independent frames belong to. The similarity measures chosen here are for simplicity during evaluation; the authors suggest that a more complex rule based system could lead to a very robust identification and separation. Our rules are equipped only to detect the sources based on their musical register with respect to each other, so one source is considered to always have a higher pitch than the other but their harmonics will of course overlap. If the sources were to intersect and switch register, this would be reflected in the output, but the outputs would still remain separated and independent. The spectral centroid [8] is a measure of the 'brightness' of a sound or as it is often referred to the 'centre of gravity' of a power spectrum. We use it here to indicate the register of the instrument. It is obtained using equation 7. The frames are normalised before the spectral centroid is obtained in order to avoid intensity ambiguities.

$$SC = \frac{\sum_{k=1}^{N} kX(k)}{\sum_{k=1}^{N} X(k)} \tag{7}$$

So we will have SC_1 and SC_2 for each output respectively. As a second indication of register we take the peak location from each of the independent short-time spectra which is assumed to be the fundamental frequency of each source at a given time. The peak is denoted as $P_1(m)$ and $P_2(m)$ for each output respectively. The next measure simply performs simple peak tracking. We define a distance measure between each of the current peaks and each of the previous peaks as in equation 8.

$$\begin{aligned} D_1 &= \left| P_1(m) - P_1(m-1) \right| \\ D_2 &= \left| P_1(m) - P_2(m-1) \right| \end{aligned} \tag{8}$$

We now have three measures which describe the randomly ordered magnitude frames. We predetermine that the source with the highest register will be contained in the vector I_1 and the source of lower register will be contained I_2. We now compare each measure to determine which source container a given independent frame should be placed in. The scores are evaluated simply as follows.

$$\begin{aligned} &L_{1++} \quad if\ SC_1 > SC_2 \quad else\ L_{2++} \\ &L_{1++} \quad if\ P_1 > P_2 \quad else\ L_{2++} \\ &L_{1++} \quad if\ D_1 < D_2 \quad else\ L_{2++} \end{aligned}$$

$$\begin{aligned} &if\ L_1 > L_2 \quad \begin{Bmatrix} I_1(k,m) = S_1'(k,m) \\ I_2(k,m) = S_2'(k,m) \end{Bmatrix} \\ &else \quad\quad\ \begin{Bmatrix} I_2(k,m) = S_1'(k,m) \\ I_1(k,m) = S_2'(k,m) \end{Bmatrix} \end{aligned} \tag{9}$$

where L_1 is the likelihood that output vector 1 from the ICA front end belongs to the source with the highest register, i.e. source 1 is assigned to container 1. The likelihood scores are compared to see which permutation of the outputs is required. At this stage we now have two independent magnitude spectrograms, one for each source.

3.4. Adaptive Filtering

There are instances where, the ICA outputs are not completely independent. This is usually evident where both sources exhibit a similar time evolution. It usually results in poor separation during those time frames. So as a final process to force more further attenuation of the unwanted source, an adaptive filtering technique is used which consists of multiplying each independent frequency frame by a normalised, smoothed and inverted version of the opposite frame. This suppresses residual effects of the unwanted sources between the frames resulting in greater separation.

$$Y(k,m) = [I_j(k,m)] \times 1 - [w(n) * I_i(k,m)] \quad i \neq j \tag{10}$$

where $w(n)$ is a suitable hanning window and $*$ denotes convolution, which is performed in order to smooth the magnitude response into something resembling an equalization curve. This curve is inverted

so that when multiplied by the opposite frame, it will suppress any residual information from the other source. This is done for each frame before resynthesis using the IFFT, equation 11.

$$y(n+mH) = w(n)\left(\frac{1}{K}\sum_{k=1}^{K} Y(k,m)\cdot e^{j\angle x_{\omega}(k,m)}\right) \tag{11}$$

where $\angle x_{\omega}(k,m)$ is the phase information from the original mixture spectrogram and *w(n)* is a suitable window of length *N*.

4. RESULTS

To evaluate the system we synthesized a mixture of bass and flute. Both sources are always active and do contain harmonic overlap. Each source is playing a melody.

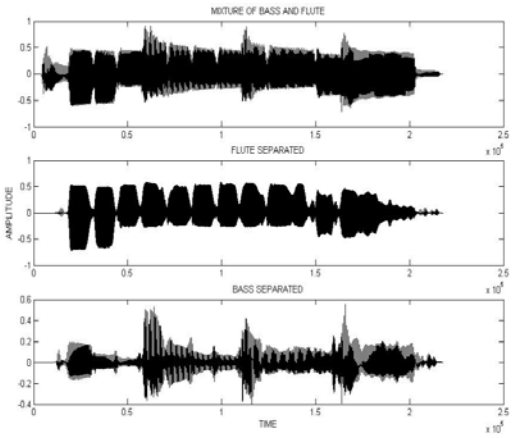

Figure 3: The top plot shows the mixture signal, the centre plot shows the resynthesised flute separation and the lower plot shows the resynthesised bass separation.

For the separations above, a 4096 point window with 75% overlap was used. The frame distance, denoted by α in equation 3 was set to 12 which corresponds to 330mS for the parameters chosen. This particular value was achieved through experimentation. A degree of separation is always achieved, but this setting provided best the results for this example.

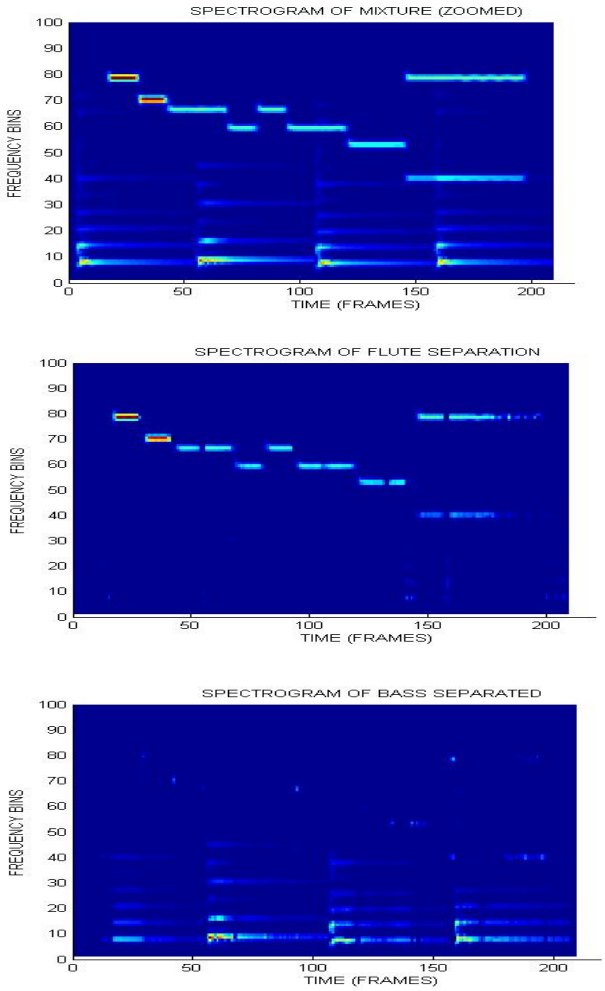

Figure 4: The spectrogram on the top shows the mixture signal followed by the flute and the bass respectively. Note that the discontinuities in the flute notes can be seen as artifacts in the bass separation.

5. CONCLUSIONS

We present a framework which is capable of single channel source separation. Although the system is currently limited by the simple source ordering rules, the results remain compelling. The method could be extended and made more robust by adding more ordering rules at the output stage. Theoretically the method could be extended to deal with more sources by taking *J* frames at the input, where *J* is the number of sources present. This makes the problem significantly more complex since there will be *J* factorial (*J*!) permutation possibilities to consider during the source ordering stage. The main limitation with the system

described is the fact that both sources must be active all of the time. To overcome this problem, a mechanism whereby two output frames can be assigned to the same source could be employed.

6. REFERENCES

[1] D 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*, June 2000

[2] 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

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

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

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

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

[7] FitzGerald, D., Lawlor, B., Coyle, E., "Independent Subspace Analysis using Locally Linear Embedding", *Proceedings of the Digital Audio Effects Conference (DAFX03),* London, pp. 13-17, 2003.

[8] Beauchamp, J. W., ''Synthesis by spectral amplitude and brightness matching of analyzed musical instrument tones,'' J. Audio Eng. Soc. 30, 396–406. 1982