# A system for providing audible separation of mixed sound sources

**Robert Bullen (1), Bjoern Erlach (2), Jonathan Abel (3)**

(1) SoundScience P/L/, Level 4, 272 Pacific Hwy, Crows Nest, NSW, Australia
(2) Centre for Computer Research in Music and Acoustics, Stanford University, CA, USA
(3) Centre for Computer Research in Music and Acoustics, Stanford University, CA, USA

## ABSTRACT

The problem of providing an audible separation of mixed sound sources is important for a number of applications, including speech recognition, noise reduction in communication channels, re-mixing of recorded music, and using environmental sound in musical compositions and film scores, as well as applications in environmental noise control. Many approaches to the problem have been investigated, each with application in a specific area. This paper presents a novel approach that would have application where: a) high-quality reproduction is desired with minimum artifacts; b) measurement using a multiple-microphone array is possible; and c) real-time performance is not required. As such it would apply particularly to audio-oriented applications, but may also have application in environmental noise. The technique involves a constrained least-squares decomposition of spectrogram values recorded at multiple microphones, together with an optional adaptive filtering step. Performance of the algorithm is described for simulated mixtures, and compared with published data for other techniques. It compares well with other systems, particularly in terms of rejection of audible artifacts.

## INTRODUCTION

Problems involving the separation of mixed sound sources arise in a number of areas, including:

- speech recognition in the presence of other talkers (e.g. Saruwatari et al 2003);
- reducing interference in communication channels (e.g. Cho & Krishnamurthy 2003) ;
- environmental noise monitoring (e.g. Bullen 2003);
- de-noising of recorded speech or music (e.g. Ellis & Weiss 2006);
- re-mixing of recorded music (e.g. Woodruff, Pardo & Dannenberg 2006); and
- using environmental sounds for music composition, film sound-tracks and other purposes.

Vincent, Fevotte and Gribonval (2003) distinguish between "audio quality oriented" applications, in which the object is to produce an audible re-creation of a source, and "significance oriented" applications in which the object is the extraction of certain features of the sound. The latter would include speech recognition, where the object is not to listen to the reconstructed speech but to determine what words are being said, and environmental monitoring, in which the object is generally to determine the level of the sound. This distinction will clearly affect the approach taken to the separation problem.

The extent of available information also has a major impact on the approach taken. Applications in which only a monophonic or stereo signal is available (e.g. Diamantaras, Petropulu & Chen 2000) will be handled very differently from applications where it can be assumed that multiple spatially-separated transducers are used for recording.

Different approaches are also generated by the use of different (assumed) properties of the sources. In particular:

- the assumption that sources are non-Gaussian and have no mutual information leads to techniques based on Independent Component Analysis (ICA - see Comonand & Juten, 2010);
- the assumption that sources are non-stationary and potentially Gaussian leads to techniques based on Non-Negative Matrix Factorization (NMF - see Ozerov & Fevotte 2010 for a recent example of this technique); and
- the assumption that sound from different sources arrives at the measurement position from different directions leads to techniques based on beamforming (e.g. Hur et al 2011).

A recent summary of approaches to the source separation problem is found in Comonand and Jutten (2010).

In addition, hybrid approaches have been investigated, examples being Sauruwatari et al (2003) and Wang, Ding & Yin (2011), in which beamforming and ICA are combined.

This paper considers a class of applications defined by the following characteristics.

- They are "audio quality oriented", and intended for re-mixing or re-using recorded sound. This has consequences in terms of toleration of distortion in the separated signals. Whereas interference from other sources (accurately reproduced) may limit the usability of the system, audible artifacts may render it completely unusable.
- Multiple microphones can be used in the recording. This effectively means the application applies to material that has been specifically recorded with a view to later separation. In fact, for cases considered in this paper the number of microphones is assumed to be at least as great as the number of sources to be extracted.

The separation techniques used in this paper are based on the direction of arrival of sound. This choice is made largely on the basis that direction-of-arrival approaches tend to achieve

a lower ratio of artifacts to interference than ICA or NMF approaches. However, the techniques differ from the traditional beamforming approach, and are based on a constrained least-squares fit to the complex spectrograms recorded at each microphone.

The orientation of the work described is toward separation of sound that is recorded externally, and for this application reverberation is not significant, but the presence of numerous interfering sources is very significant. This orientation drives the evaluation procedures adopted, although a comparison with other techniques is made in the context of reverberant environments.

## SEPARATION ALGORITHMS

### Problem Formulation

The sound pressure measured at M microphones can be represented as M real-valued signals $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_M(t))^T$. These are to be expressed as a liner mixture of N source signals $\mathbf{s}(t) = (s_1(t), s_2(t), \ldots, s_N(t))^T$, which may be differentially filtered before reaching the microphones, plus additive noise. In a room, the filters will represent the room impulse response between each source and each microphone. For anechoically-recorded sound, the filters will represent pure delays.

Hence

$$\mathbf{x} = \mathbf{a} * \mathbf{s} + \mathbf{u} \quad (1)$$

where **a** is an unknown M x N matrix of filters, * represents convolution with the matrix elements, and **u** is an unknown M x 1 vector of noise signals which are independent of each other and of the source signals.

In a physically realistic measurement there will be multiple noise sources contributing to **x**, some of which will be identifiable and/or of interest, while others are not. **s** can be considered to represent the N sources that are of interest, while **u** includes sound from all other sources, which will generally be much more numerous. In this way, an underdetermined problem (N > M) can be re-cast as an over-determined problem (N < M) with a significant noise component.

In addition, where reverberation is present this can often be satisfactorily modelled as direct sound plus a noise-like reverberant component that is independent of the direct signal. If this reverberation is also included in **u**, the problem can be reduced to the case where all filters in **a** are pure delays.

Taking a Short-Term Fourier Transform (STFT) of (1) transforms the convolutive filters in **a** into complex multiplications in the frequency domain:

$$\mathbf{X}_k(\omega) = \mathbf{A}(\omega)\, \mathbf{S}_k(\omega) + \mathbf{U}_k(\omega) \quad (2)$$

where **X**, **A**, **S** and **U** represent the (complex) discrete Fourier transforms of **x**, **a**, **s** and **u** respectively, $\omega = 2\pi f$ where f is the frequency of a bin, and k indexes the sample frame. If the filters in **a** are pure delays, then

$$\mathbf{A}_{ij}(\omega) = \exp(-j\omega\, t_{ij}) \quad (3)$$

($j = \sqrt{-1}$). Here $t_{ij}$ is the delay for signal j between the origin of co-ordinates (where signal j is defined) and microphone i. Hence

$$t_{ij} = -\, \mathbf{m}_i \,.\, \mathbf{d}_j / c \quad (4)$$

where $\mathbf{m}_i$ is the vector from the origin to microphone i, $\mathbf{d}_j$ is a unit vector in the direction of source j, and c is the speed of sound.

### Separation with Known Direction of Arrival

If the number of sources, N, and the direction of each source, $\mathbf{d}_j$, are assumed to be known, then $\mathbf{A}(\omega)$ is known and (2) requires selecting $\mathbf{S}_k(\omega)$ and $\mathbf{U}_k(\omega)$ for each $\omega$ to partition $\mathbf{X}_k(\omega)$ between $\mathbf{A}(\omega)\mathbf{S}_k(\omega)$ and $\mathbf{U}_k(\omega)$.

(From this point we remove the explicit dependency on k and $\omega$, but understand that all variables refer to a single time-frequency point in a STFT.)

One option to provide the separation in (2) is to minimize **U** in the least-squares sense (i.e. minimise $\mathbf{U}^H\mathbf{U}$), so that **S** is given by the standard regression formula

$$\mathbf{S} = (\mathbf{A}^H\mathbf{A})^{-1}\, \mathbf{A}^H\, \mathbf{X} \quad (5)$$

where H represents hermitian transpose.

However, the matrix $\mathbf{A}^H\mathbf{A}$ will often be ill-conditioned, particularly at low frequencies where all elements of **A** are close to 1, or where two sources are on almost opposite sides of the origin so that $\mathbf{d}_1 \sim -\, \mathbf{d}_2$. This leads to unstable solutions including out-of-phase sources with very high power.

A common way to avoid such ill-conditioned matrices, and in general to increase the smoothness of least-squares solutions, is to replace (5) with

$$\mathbf{S} = (\mathbf{A}^H\mathbf{A} + \lambda\mathbf{I})^{-1}\, \mathbf{A}^H\, \mathbf{X} \quad (6)$$

where $\lambda$ is a constant, usually taken to be small, and **I** is the NxN identity matrix. This is sometimes known as Tikhonov regularisation (Press et al, 2007).

Using the notation $\mathbf{S}^H\mathbf{S} = ||\mathbf{S}||$, solving (6) is equivalent to solving (2) with the additional constraint that $||\mathbf{S}|| = \alpha$ where $\alpha$ is a pre-determined constant - that is, a quadratically constrained least-squares fit to the data. As $\lambda$ increases, the size of the solution, $||\mathbf{S}||$, decreases from its unconstrained value ($\lambda = 0$).

A physically motivated choice for $\alpha$ is the constraint that at each time-frequency point, the total power in all sources should not exceed the mean power measured by the microphones -

$$||\mathbf{S}|| <= (1/M)\, ||\mathbf{X}|| \quad (7)$$

Solutions of (6) can be found, using efficient procedures, with values of $\lambda$ increasing from zero until (7) is satisfied. This then gives a vector **S** of complex-valued Fourier coefficients at a specific time-frequency point. The procedure is repeated at all frequency points and in all sample frames, and each source signal is reconstructed from the co-efficients by inverse transformation and overlap-add between frames.

### Determination of Direction of Arrival

For a given number of sources in given directions, the procedure above estimates the complex-valued STFT coefficients $X_k(\omega)$ at each microphone, for each time-frequency point. If these estimates are labelled $\hat{X}_k(\omega)$, the overall goodness of fit can be estimated by

$$C_1 = \sum_{\omega,k} \| X_k(\omega) - \hat{X}_k(\omega) \| \quad (8)$$

To avoid biasing the result toward time-frequency points with high power, the estimate can be normalized:

$$C_2 = \sum_{\omega,k} \| X_k(\omega) - \hat{X}_k(\omega) \| / \| X_k(\omega) \| \quad (9)$$

$C_2$ can be taken as a cost function and minimized over the space of numbers of sources and arrival directions, using any appropriate minimisation technique. If it is possible that sources could be moving, then the sum over sample frames k should be limited to an appropriate time scale. In the simulations below, although sources are not moving, source directions are estimated independently in each 1-second "chunk".

To reduce the time and complexity of this calculation, the simplification is made in the remainder of this paper that all sources lie in a plane, which can be identified with the horizontal. This will generally be true for externally-recorded sound, and often also for internally-recorded sound. Hence the arrival direction can be described by a single angle.

Even with this simplification, minimisation of $C_2$ using values of $\hat{X}_k(\omega)$ calculated at all time-frequency points would be very computationally expensive. However, it is not necessary to calculate at all points – trials indicate that a relatively small number of frequency points can be used to distinguish between arrival directions. In simulations below, eight frequency points are used from each STFT – that is, the summation in (9) is restricted to eight values of ω. If there is *a priori* knowledge of the likely frequency content of sources, the frequencies used can be chosen correspondingly, to improve noise rejection.

In simulations below, the number of sources to be found in each 1-second sample is determined by the user. It corresponds to the number of simultaneous sources that are considered interesting and/or that can be successfully detected. For the simulations described below, good angle determination appears to be possible for up to about 4 simultaneous sources.

A standard Nelder-Mead simplex algorithm is used to minimise $C_2$ over possible arrival directions for a given number of sources.

### Finding Connected Sources

Once a set of source angles has been identified for each 1-second "chunk" of data, these can be joined to form "connected sources" spanning longer time periods. The algorithm for this process is based on a "link strength" defined between each source angle and every other source angle in all chunks. The link strength is based on:

- the difference between the source angles in the two chunks;
- the time difference between chunks (in simulations, it is always zero for chunks separated by more than 2 secs); and
- the similarity between the recovered spectra at the two source angles.

The algorithm is similar to standard edge-detection algorithms used in image processing, and proceeds as follows (see Figure 1).

1. Select the pair of un-linked sources with the highest link strength (provided it is greater than some starting criterion).
2. Join the right-hand source to the source on the right with the highest link strength, and repeat, until the link strength is below a stopping criterion or the end of the data is reached. Similarly for the source on the left.
3. Repeat from 1.

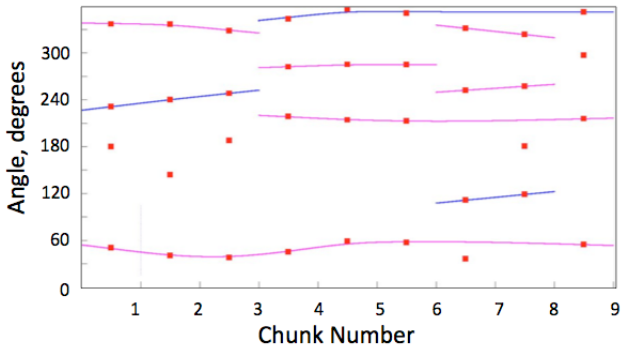


**Figure 1** Illustation of selection of sources in each chunk (in this case four) and joining to form connected sources

A set of points that has been linked across chunks in this way is referred to as a "connected source". Figure 1 shows nine connected sources formed from separations performed in nine chunks, with four source angles detected per chunk. Connected sources may have slowly-changing directions and may or may not span the time range of the data. A cubic spline fit to the source angles in a connected source (allowing for wrapping at 360 degrees) then provides an estimated source angle at any time.

Now, for each sample frame in the original data, the source angles for active connected sources can be used in the procedure described above to estimate STFT coefficients at each frequency. Hence, estimated source signals can be reconstructed via inverse transformation and overlap-add.

### Post-Separation Filtering

It is possible to include an adaptive filter after the separation process above, to further reduce the residual noise component. This is based on including an additional source in the final separation, located as far as possible in angle from any of the detected sources. This is intended to represent a pure noise component.

The filter reduces the level of detected sources at time-frequency points where the "source" level is similar to that in the "noise" component, and leaves it unchanged where the "source" level is much higher. In practice the filter is a 200-lag FIR filter with continuously-varying coefficients based on levels averaged over bark intervals and over several frames. An attempt is made to locate large changes in the coefficients at transients in the signal, to reduce audible "pumping".

Because this filter is non-linear, it inevitably increases the level of target distortion in the reproduced signal, while reducing the interference from noise. The optimal trade-off between these will depend on the signal-to-noise ratio in the original, and on the purpose of the separation.

## TEST PROCEDURES

### Test Signals and Orientation

Test signals consisted of up to four "foreground" sound sources:

- Male voice;
- Female voice;
- Jackhammers; and
- Bus,

while "background" sounds were traffic noise or noise from a crowded café.

All files were 44,100 Hz, 16 bit, mono with 10 seconds duration. Foreground sources were scaled in level, assigned an arrival angle and delayed appropriately for each microphone position in an assumed array.

For "background" sources, 30 independent samples from the same file were assigned to 30 random angles, rendered at each microphone position and added, giving an approximation to a diffuse noise field. Because results were sometimes dependent on the "background" angles selected, in each simulation five separate runs were performed with different random angles assigned to the 30 "background" signals. Quoted results represent mean values over these five runs.

Microphone configurations tested were horizontal circular arrays of 5, 10 and 20 microphones, with radius 5, 10 and 50 cm.

The simulated total signals at all microphones were passed to the separation algorithm described above. Four source angles were sought in each 1-second "chunk", and these were then joined as described above to form connected sources. STFTs used 8192-sample frames with 50% overlap.

In all cases except as specifically noted below, the separation procedure spontaneously produced connected sources at angles within 5° of each of the actual sources, for at least part of the 10 secs duration of the signal. These connected sources were considered to be estimates of the relevant actual sources. Where no source was found for part of the time, the estimated source signal was set to zero in that period.

### Evaluation

Evaluation of results was performed using software described in Emiya et al (2011). That paper describes a number of subjective and objective methods for rating the quality of audio source separation. Evaluations in the present paper are based on the objective parameters defined in Emiya et al. These are derived using the difference between the (known) true target signal and the reconstructed signal. The ratio of signal power to the power in this difference is the Signal to Distortion Ratio (SDR). The difference is then partitioned into:

- a component associated with delayed versions of the target ("target distortion");
- a component associated with other known signals ("interference"); and
- the remaining difference ("artifacts").

The relative importance of each of these is described by a power ratio, as described in Emiya et al.

In the simulation tests described here, each of the "foreground" sources in turn was assigned as the target signal. The other foreground sources and the total "background" signal, evaluated at the origin, were all considered as interfering sources. The artifacts component represents distortion introduced by the decomposition process itself.

## TEST RESULTS

### "Baseline" Tests

The "baseline" test orientation consisted of five microphones in a circular array 0.1m in radius. The four "foreground" sources were positioned at angles of 36° (male voice), 127° (female voice), 202° (jackhammer) and 284° (bus). The total (unweighted $L_{eq}$) levels of all sources were equal. Figure 2 shows the time-waveform of each source. Background noise was traffic, and the total background level was +10dB, 0dB or -10dB relative to the four source levels.

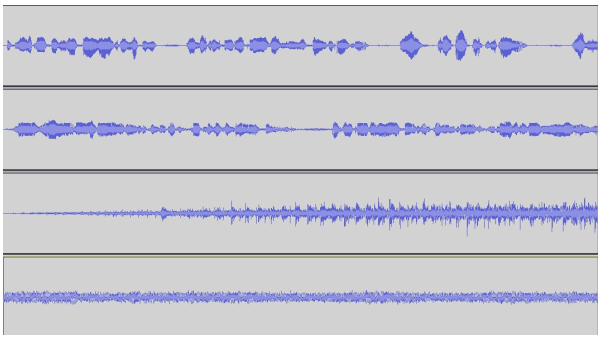

**Figure 2** Time waveforms of test signals - female voice, male voice, jackhammer and bus

Figure 3 shows the SDR for each of the recovered sources (without post-separation filtering). Where the microphone signal is dominated by background (10dB higher than the source levels), each of the sources can be recovered with an SDR of about -3dB. Where background is 10dB below the source levels, the SDR for the recovered sources is better than 10dB.

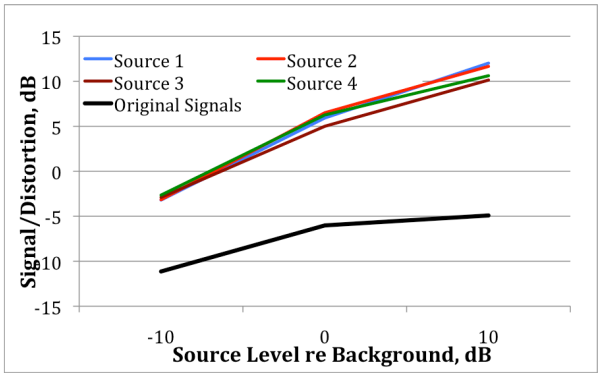


**Figure 3** Signal / Distortion ratios for each of the recovered sources in a mixture of four sources and background

SDRs are quite similar for each of the test sources. They are slightly lower for the jackhammer source, due to the fact that its level varies significantly over the 10 sec interval, and in some runs the source was not detected for part of the time. (Note that in one of the five runs at -10 dB, the bus source

was not detected at all. SDRs for this case are averaged over the remaining runs.)

Figure 4 shows the individual components of the total distortion. Notably, the largest component is from interference, rather than target distortion or artifacts. This is significant for applications where the object is to reproduce the recorded source accurately, and some "bleed" of other sound (accurately reproduced) may be acceptable.

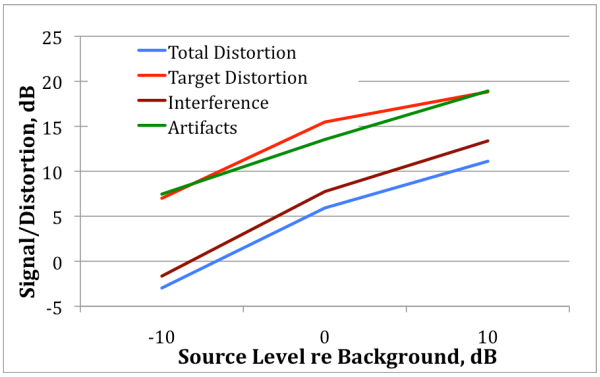


**Figure 4** Components of the signal/distortion ratios shown in Figure 1 (average over all sources)

Figure 5 shows a comparison between SDRs with and without the post-separation filtering step described above. Filtering has the effect of improving the SDR by about 4dB in high-background situations, but reduces it by about 2dB in low-background situations. This is because filtering improves the rejection of interference, but at the expense of target distortion.

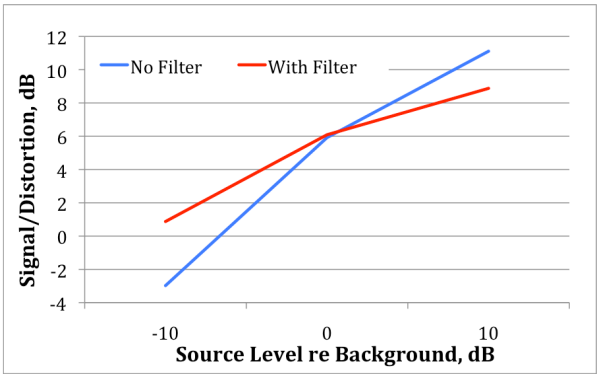


**Figure 5** Signal/distortion ratios with and without a post-filtering step (average over all sources)

## Changing the Number of Detected Sources

Figure 6 shows the performance when fewer than four sources are present in the mixture, with all sources at 0dB re the background. Surprisingly, with fewer sources to detect, the overall detection performance does not improve, and if anything appears to slightly decrease. However, with fewer sources the levels of target distortion and artifacts appear to reduce.

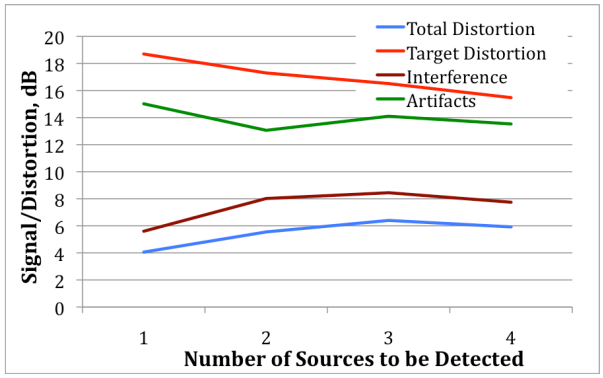


**Figure 6** Signal/distortion ratio for separation when fewer sources are present in the mixture (average over all sources in the mixture)

## Changing Number of Microphones

Again surprisingly, Figure 7 indicates there is no improvement in separation when more microphones are added to the 0.1m radius circle. (The tests shown used four sources with background at 0dB re the sources.) This is in contrast to beam-forming techniques in which adding further transducers increases performance. However, with beam-forming a much larger number of microphones than the minimum of five used here is required to produce acceptable sensitivity.

It should be noted that the above results use simulated signals, and hence do not include errors due to imprecision in microphone locations and other hardware-specific sources. When these are included, it is likely that performance would improve somewhat as the number of microphones increases.

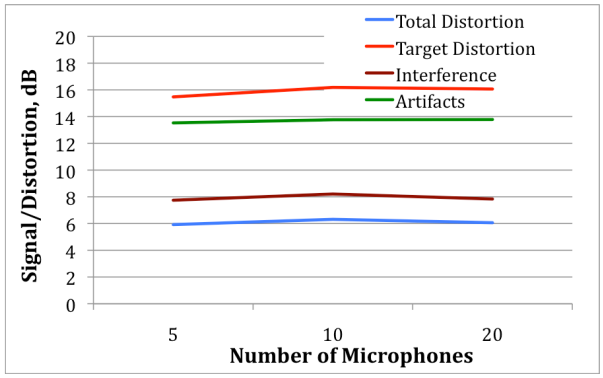


**Figure 7** Signal/distortion ratio for separation with different numbers of microphones (average over all sources)

## Changing The Size of Microphone Array

Figure 8 shows the effect of changing the radius of the circle of five microphones (with four sources, background at 0dB relative to sources). Reducing the radius below 0.1m has very little effect, while increasing it to 0.5m results in significantly reduced SDR, as well as a proportional increase in target distortion.

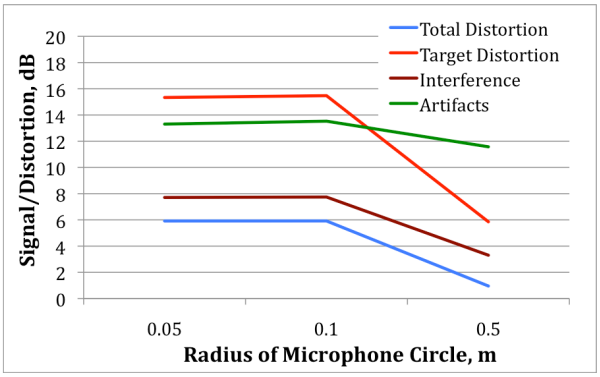


**Figure 8** Signal/distortion ratio for separation with different radii of the microphone array (average over all sources)

### Changing the Spectrum of Background Noise

The traffic background noise used in the above testing has a spectrum similar to the bus source, but significantly lower in frequency than the other sources. Figure 9 shows the separation performance when noise from a café is used as the background, at various levels re the four sources, with and without a post-filtering step. It is clear that performance with the café background is lower than with the traffic background, by about 2dB.

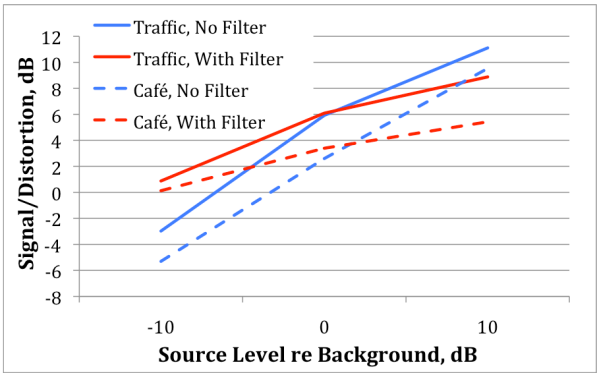


**Figure 9** Total signal/distortion ratio with cafe background noise vs traffic (average over all sources)

### Comparison With Other Results

Of the large number of source separation studies performed over the last ten years, few have considered the class of problems considered in this paper. In addition, many have used incompatible units in reporting results. For example, "signal to noise ratio" is often quoted in terms of a ratio of spectral powers, ignoring phase distortion.

The most comparable data set to these results is a section of the 2011 Signal Separation Evaluation Campaign (SiSEC 2011), described in Araki et al (2012). (A similar campaign was conducted in 2010, but in that case the available data are not sufficient to perform a full comparison with results using the present techniques.) The campaign involved several teams attempting to separate sources from a number of types of mixture, with the results reported using software very similar to that used in the above analysis. The most relevant of these is a set of recordings of speech in real-world environments – a café, a subway and a square. Unfortunately, results are available only for stereo recordings, so the techniques described above can find at most two distinct sources in addition to a general background. For recordings in the square, there are a number of other distinct sources – vehicles, other people, etc., and this means that with only two microphones the above techniques do not always find the speech source. For the other two environments, however, the reverberant background means that the speech can be clearly separated.

Figure 10 shows a comparison of the SDR found using constrained least-squares decomposition with results reported from the three study teams undertaking this task. Two of these used variants of ICA (described in Nesta and Matassoni, 2011) while the third used a form of sparse matrix decomposition related to NMF (described in Ma et al, 2010). As expected, the reverberant environment of the subway makes separation more difficult for all techniques. Nevertheless it is clear that the overall SDR for the constrained least-squares analysis is similar to the other techniques.

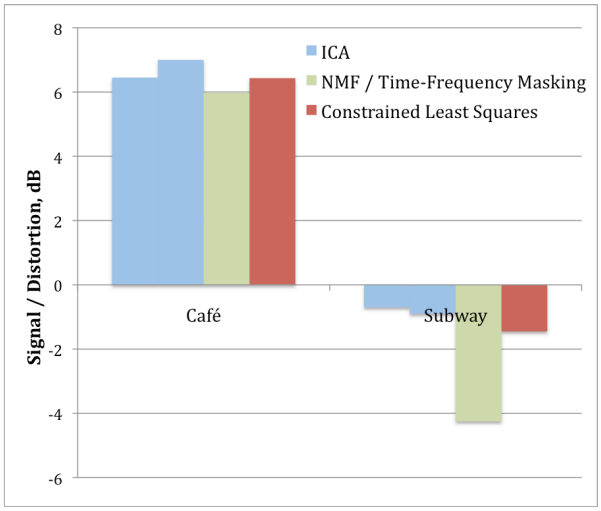


**Figure 10** Total signal/distortion ratio for separation of a single speech source from recorded background

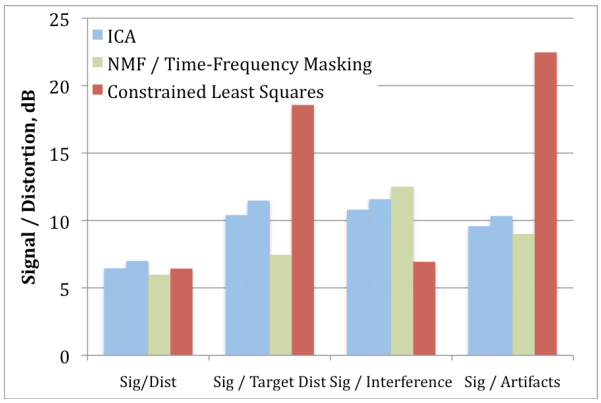


**Figure 11** Components of the signal/distortion ratio for separation of a single speech source from recorded background - cafe

However, Figure 11 shows a breakdown of the sources of distortion (for the café scenario). While the other techniques produce significant levels of both target distortion and artifacts, constrained least-squares produces significantly less of these effects, with distortion being largely due to interfer-

ence. As noted above, in some applications this difference would be quite significant.

## CONCLUSION

A novel technique is proposed for separation of sound sources in circumstances where the quality of the separated audio is critical. It involves the use of multiple microphones - at least one microphone per source of interest – and sources must be at a specific angle to the measurement point. Any other sources, as well as reverberation, are considered as distributed "background" noise.

The proposed technique involves a constrained least-squares fit to each set of the complex time-frequency values in the recorded spectrograms. Details of the performance of the technique under various conditions are described in the body of this paper.

Although direct comparison with other techniques is so far limited, the constrained least-squares method appears to provide a similar overall signal-to-distortion ratio to other state-of-the-art techniques, but with significantly lower levels of target distortion and artifacts.

The technique could have application in audio recording and post-processing, as well as identification of recorded speech and environmental sounds.

## REFERENCES


Araki S., Nesta F., Vincent E., Koldovsky Z., Nolte G., Ziehe A. & Benichoux A, 2012. 'The 2011 Signal Separation Evaluation Campaign (SiSEC2011): Audio source separation' *Proceedings of 10th International Conference on Latent Variable Analysis and Signal Separation*, pp 414-422.

Bullen R., 2003, 'Long-Term Environmental Monitoring and Noise Source Identification', *Acoustics Australia* vol 31, no 1, pp 23-27.

Cho, J. & Krishnamurthy, A. 2003, 'Speech enhancement using microphone array in moving vehicle environment', *Proceedings of IEEE Intelligent Vehicles Symposium*, pp 366-371.

Comonand P. & Jutten C. 2010, *Handbook of Blind Source Separation*. Academic Press.

Diamantaras K.I., Petropulu A.P. & Chen B. 2000, 'Blind two-input-two-output FIR channel identification based on second-order statistics' *IEEE Transactions on Signal Processing* vol 48, pp 534-542

Ellis D.P.W. & and Weiss R.J. 2006, 'Model-based monaural source separation using a vector-quantized phase-vocoder representation' *Proceedings of International Conference on Acoustics, Speech and Signal Processing 2006* vol 5.

Emiya V., Vincent E., Harlander N. & Holman V. 2011, 'Subjective and objective quality assessment of audio source separation'. *IEEE Transactions on Audio, Speech and Language Processing* vol 9, no 7, pp 2046-2057.

Hur Y., Abel J., Park Y.-C. & Youn D. 2011. 'Techniques for synthetic reconfiguration of microphone arrays' *J. Audio Eng. Soc*. vol 59, no 6, pp 404-418.

Ma W., Yu M., Xin J. & Osher S. 2010. 'Reducing musical noise in blind source separation by time-domain sparse filters and split bregman method' *Proceedings of Interspeech 2010*, pp 402-405.

Nesta F. & Matassoin M. 2011. 'Robust automatic speech recognition through on-line semi-blind source extraction' *CHIME Workshop 2011, Florence, Italy*

Ozerov A. & Fevotte C 2010, 'Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation' *IEEE Transactions on Audio, Speech, and Language Processing* vol 18 no 3, pp 550-563.

Press W., Teukolsky S., Vetterling W. & Flannery B., 2007. *'Numerical Recipes: The Art of Scientific Computing'* Cambridge University Press. (Section 19.5)

Saruwatari H., Kurita S., Takeda K., Itakura F., Nishikawa T. & Shikano K., 2003 'Blind source separation combining Independent Component Analysis and beamforming', *EURASIP Journal on Applied Signal Processing*, vol 11, pp 1135-1146.

Vincent E, Fevotte C & Gribonval R, 2003 'A tentative typology of audio source separation tasks' *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Source Separation*, pp 715-720.

Wang L., Ding H & Yin F. 2011, 'Target speech extraction in cocktail party by combining beamforming and blind source separation' *Acoustics Australia* vol 39, pp 64-68.

Woodruff J., Pardo B. & Dannenberg R. 2006, 'Remixing stereo music with score-informed source separation' *Proceedings of 7th International Conference on Music Information Retrieval.* pp 314-319.