2008-01-01

Single Channel Sound Source Separation Combining Delay Estimation and the AdRess Algorithm

Mark Leddy
Dublin Institute of Technology

Dan Barry
Dublin Institute of Technology, dan.barry@dit.ie

David Dorran
Dublin Institute of Technology, david.dorran@dit.ie

Eugene Coyle
Dublin Institute of Technology, Eugene.Coyle@dit.ie

Recommended Citation

Leddy, M., Barry, D., Dorran, D. & Coyle, E. Single channel sound source separation combining delay estimation and the AdRess algorithm. Paper given at the *Irish Signals and Systems Conference, National University of Galway, 2008.*
http://www.audioresearchgroup.com/

Single Channel Sound Source Separation combining Delay Estimation and the ADRess algorithm

Mark Leddy, Dan Barry, David Dorran and Eugene Coyle.

Audio Research Group
Dublin Institute of Technology, Kevin St.,
IRELAND

E-mail: `mark.leddy@dit.ie` `dan.barry@dit.ie`
`david.dorran@dit.ie` `eugene.coyle@dit.ie`

Abstract — **A method for single channel source separation is proposed in this paper, which uses estimates of the delay co-efficient of individual sources within an echoic mixture using autocorrelation, following which a "pseudo-stereo mixture" is generated, to which the ADRess algorithm can be applied. The system is evaluated in a theoretical situation, where the mixture signal to be separated consists of two individual source signals, and a delayed version of each signal. Estimates of the individual delay lengths are made and then used to create a pseudo stereo mix, where one channel consists of the original mixture signal, and the second channel consists the original mixture signal shifted by the length of the delay calculated for each source. The ADRess algorithm is then used to separate sources from the new pseudo stereo mixture.**

Keywords — **Single channel, Blind Source Separation, ADRess, BSS**

I Introduction

Sound source separation is the process of separating individual source signals from a mixture of multiple sources. Techniques such as DUET [3], and ADRess [2] are used to perform source separation, however they require the use of at least two different mixtures of the same sources. In the case of audio separation, two microphones or a two channel recording are required. These techniques have been shown to work well and can produce robust and high-quality results.

Factorisation based approaches such as PCA [7], ICA [6] and NMF [4] have been applied using just a single mixture of sources. These techniques have shown their effectiveness for certain tasks such as musical transcription. However, in general, the output resynthesis quality of the separations are not as robust as the above 2-channel techniques.

Proposed here is a source separation technique, which creates a two-channel, pseudo-stereo mixture from a single-channel mixture signal. From this stereo mixture the ADRess algorithm can be employed to separate a single source from the mixture. By using this technique it is shown that robust separation results from just a single channel mixture signal is possible.

The aim of this paper is to show that the described technique will work for simplified, synthetic cases, in order to pave the way for further exploration.

II Delay Model

The theoretical model, used to represent a single source in an echoic environment, is illustrated in Fig. 1. Here S represents a source signal that is transmitted. Acoustic pressure waves will travel along a direct path, to the sensor, and also reflected paths, rebounding off surfaces, before accumulating at the target sensor x. This mixture model is represented in equation (1) [9].

$$x(t) = \sum_{i}^{N} \alpha_i s(t + \Delta t_i) \qquad (1)$$

where $\alpha_i \geq 0$ represents the attenuation over the extra distances travelled along each reflected path

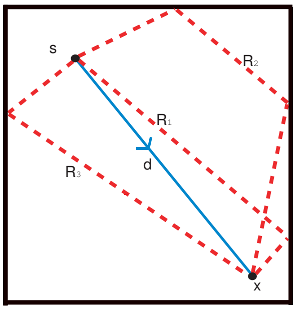


Fig. 1: Figure shows the direct path d, from the source s, to the sensor x. The three reflected paths R_i are also shown. Due to the extra distance traversed to reach the sensor, each R_i will be attenuated compared to the direct path. Similarly, due to the extra time taken to travel the the reflected paths, upon reaching the sensor they will appear as delayed versions of the source.

R_i, in comparison to the direct path d, as illustrated in Fig. 1. R_i is the path travelled by signals to reach the sensor along the i^{th} reflected path, and d is the path taken by the signal, to reach the sensor along the direct path. The resulting delay due to the extra time required to traverse the reflected path, compared to that of the direct path, $(R_i - d)$, is represented by Δt_i.

In reality, an infinite number of possible reflections will be present in an echoic environment. Typically however, many environments will have a small number of strong reflections. This is illustrated by measuring the impulse response of an echoic environment, see Fig. 2. However, for the purposes of this paper, a simplified situation will be demonstrated.

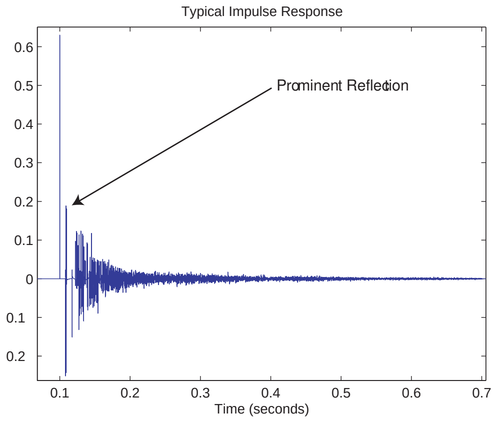


Fig. 2: Typical Impulse Response of an echoic environment. A prominent reflection can clearly be distinguished. If these reflections can be estimated, separation may be possible using the technique described in this paper.

The simplified situation presented in this paper assumes that only two sources are present, and that each source is only reflected once. This situation is presented in Fig. 3, the model described in equation (2) is used.

$$\begin{aligned} x(t) = [s_1(t) + \alpha s_1(t + \Delta t_1)] + ... \\ +[s_2(t) + \beta s_2(t + \Delta t_2)] \end{aligned} \tag{2}$$

where $s_i(t)$ represents sources received by the sensor at time t. The value Δt_i represents the extra time taken for a reflected signal to reach the sensor. The attenuation coefficients of the first and second signals, having travelled the extra reflected distance before reaching the sensor, are represented by α and β respectively. Hence the sensor mixture will consist of each source, and one delayed and attenuated version of each source.

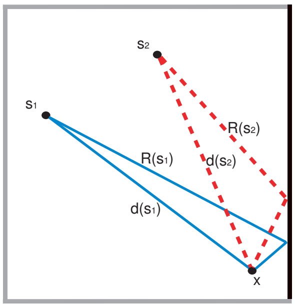


Fig. 3: Theoretical model used to illustrate the situation in which source separation will be performed. The system contains 2 sources s_i. Each will take a direct path, $d(s_i)$ to the sensor x, and also a reflected path $R(s_i)$.

III Auto-Correlation and Stereo Alignment

Autocorrelation, is used estimate the delay coefficients Δt_i present in the mixture signal $x(t)$, equation (3).

$$R_m(k) = \sum_{j=0}^{N-1} x(t + j)x(t) \tag{3}$$

where N is the length of the mixture signal $x(t)$. Autocorrelation can be used to find periodic or repeating patterns within the signal. Applying autocorrelation to the mixture signal will give an estimation of the length of the delays Δt_1 and Δt_2, see Fig. 4. The autocorrelation function here shows two significant peaks which indicate time delays Δt_1 and Δt_2. It is difficult to distinguish which delay belongs to which source at this point.

Once the delay coefficient has been established a second channel is created. This will consist of the original mixture signal, shifted forward in time by the estimated delay coefficient Δt_i. The delayed and attenuated version of the i^{th} source, will then be time-aligned with the 'original' source within the mixture. This results in a two channel mixture, consisting of $x(t)$ and $x(t - \Delta t_i)$, which will be used to recover the i^{th} source.

In theory, it should be possible to separate single sources, from a mixture of a large number of

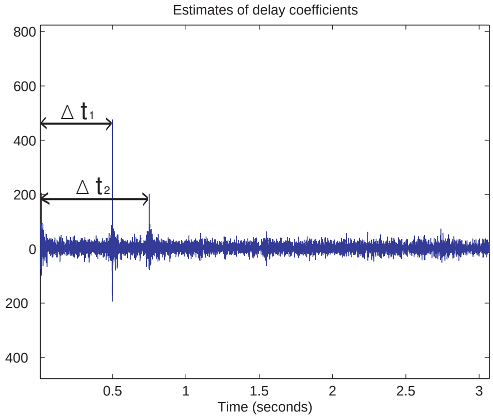


Fig. 4: Autocorrelation of the mixture signals. The peaks are used to estimate the delay coefficients Δt_i.

sources. What limits this, is the increased difficulty associated with measuring the delay coefficients using autocorrelation. As long as accurate estimates can be found, such as those in Fig. 4, it should be possible to perform separation for n sources. However as n increases, time-frequency overlap will reduce the resynthesis quality attainable with ADRess.

IV Stereo Space Source Separation

Having aligned the mixture signals into a pseudo-stereo, two-channel mixture, the ADRess algorithm can be used to separate the desired source signal. Taking a stereo mixture, the ADRess algorithm separates sources according to their lateral displacement within a stereo field. A stereo localisation, or lateral displacement effect, occurs when there is a difference in the intensity of a single source between each channel. This intensity difference allows for the creation of a histogram representing the stereo space, where by the position of sources within the stereo field can be located.

In order for ADRess to operate, the linear intensity mixing model must apply. Essentially; the sources for separation in the left and right mixture must be phase coherent, ie. time aligned. The lateral displacement must only be a function of intensity difference between each channel. The time alignment procedure and pseudo stereo mixture creation, attempts to satisfy these criteria. For further information about ADRess, see [2].

By taking our mixture signal $x(t)$ as one channel, and $x(t - \Delta t_i)$ as a second, the ADRess algorithm can then be applied. Our source $s_i(t)$, and its delayed version $\alpha s_i(t + \Delta t_i)$, have now been aligned in time. If the attenuation coefficient is negligible, ie $\alpha = 1$, then the source i will have the same intensity in both channels, and hence will be located in the center of the stereo field. Generally the attenuation coefficient of the delayed source is < 1, and this will cause the source to be located in the left half of the stereo energy histogram.

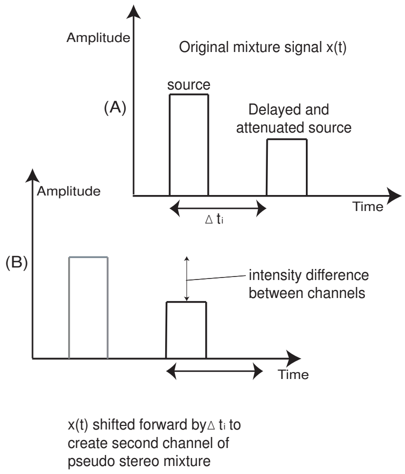
The ADRess algorithm allows for real-time plotting of this histogram. This permits the localisation of the source, and allows the user to choose the correct attenuation factor manually, as indicated by a peak on the histogram.

V Testing

In order to test the systems effectiveness, speech signals were used rather than musical signals. Speech signals display approximate W-disjoint orthogonality [5], which means the time-frequency representations of the signals do not excessively overlap.

Conversely, it is the nature of musical signals to harmonically overlap. For this reason multiple sources may contribute to a single time frequency point. Also pitched musical notes, for example a note played on a piano, will often tend to last longer than an utterance of speech.

These attributes of musical signals may make it more difficult to get an accurate delay estimate using auto-correlation. Also the stationary length of musical signals, or the amount of time a note persists, will typically be longer than the delay coefficient of the first reflection see Fig. 6. If this is the case, the musical sound, and its delayed equivalent will be added together, causing an increase in the magnitude where they overlap. This new magnitude will cause the intensity to vary erratically, essentially causing the attenuation estimate to vary also. The stationary length of speech is usually less than that of musical signals see Fig. 5, and



Amplitude
Original mixture signal x(t)
source
(A)
Delayed and attenuated source
Amplitude
Time
△ tᵢ
(B)
intensity difference between channels
Time
x(t) shifted forward by△ tᵢ to create second channel of pseudo stereo mixture

Fig. 5: Illustrates the situation when the stationary length of a signal is $< \Delta t_i$, the intensity difference will be constant. This allows robust separation using the ADRess algorithm. (A) denotes the original mixture signal $x(t)$. (B) represents the newly created channel, $x(t - \Delta t_i)$, consisting of the original mixture signal, shifted by the estimated delay.

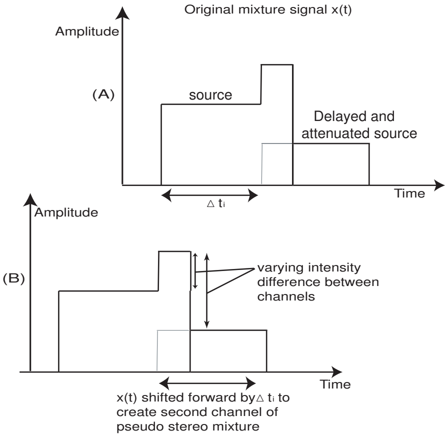


Fig. 6: Illustrates that when the stationary length of a signal is $> \Delta t_i$, the intensity difference may vary. This will mean that the i^{th} signal will be spread over a large region of the stereo space, resulting in poorer separation with the ADRess algorithm. (A) denotes the original mixture signal $x(t)$. (B) represents the newly created channel, $x(t - \Delta t_i)$, consisting of the original mixture signal, shifted by the estimated delay.

this technique is developed with speech separation in mind. For illustrative purposes, two test signals are used in this paper. Each is approximately 7 seconds of male and female speech respectively, see Fig. 8.

VI Results

The test signal used here for illustration, is a single channel mixture signal, of the form described by equation (2), see Fig. 7. The system was tested by employing 50 linearly spaced delay values, between 0 and 500milliseconds, and testing the systems ability to separate the male voice from the mixture described.

As a performance measurement, the signal to interference ratio (SIR), suggested in [8] is used. This measure was chosen as the model will contain no noise, and the only error in the estimation of the source signal $s_i(t)$, will be the contribution of the other source $s_j(t)$, and the reflected signals $s_i(t + \Delta t_i)$ and $s_j(t + \Delta t_j)$, $\forall i \neq j$. The SIR is determined both pre-separation, equation (4), and post-separation, equation (5). Their respective differences are then used to indicate the level of noise rejection achieved using the technique described here.

$$\mathrm{SIR}_{\mathbf{pre}} := 20 \log_{10} \frac{||S_{\mathbf{target}}||}{||S_{\mathbf{mixture}} - S_{\mathbf{target}}||} \tag{4}$$

$$\mathrm{SIR}_{\mathbf{post}} := 20 \log_{10} \frac{||S_{\mathbf{estimate}}||}{||S_{\mathbf{estimate}} - S_{\mathbf{target}}||} \tag{5}$$

where $S_{\mathbf{target}}$ is the test signal to be separated, $S_{\mathbf{estimate}}$ represents the separated estimation of the target signal. $S_{\mathbf{mixture}}$ is the mixture of signals represented by equation (2).

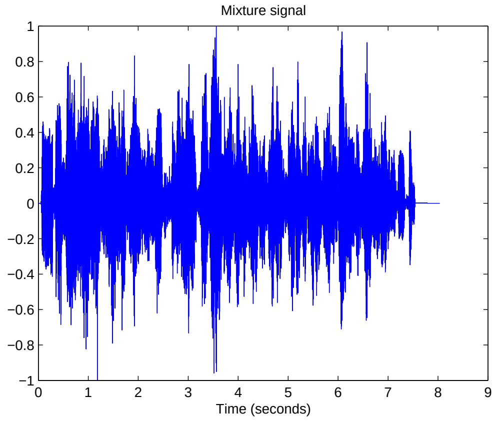


Fig. 7: Mixture signal consisting of male speech sample, female speech sample, and an attenuated and delayed version of each.

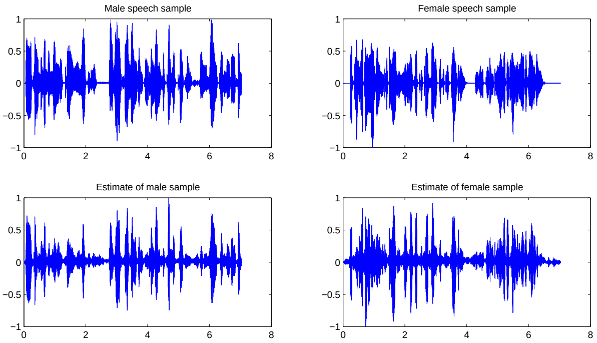


Fig. 8: Male(top left), and Female(top right), Male and Female estimates (bottom left and right respectively).

Prior to separation, the mixture signal was generated such that the interfering sources are $4dB$ louder than the source of interest, leading to a signal to noise ratio of $-4dB$. After applying the process described in this paper, analysis shows that an average of $+4dB$ of signal to interference ratio has been achieved, resulting in an average signal to noise ratio increase of $8dB$.

$\mathrm{SIR}_{\mathbf{post}}$ for a delay coefficient from 0.5 to 0.1 seconds averages about $+4dB$. For this delay timeframe, an approximate $+8dB$ noise rejection difference is maintained over $\mathrm{SIR}_{\mathbf{pre}}$. As the delay coefficient decreases below 0.1 seconds $\mathrm{SIR}_{\mathbf{post}}$ diminishes. However, even though $\mathrm{SIR}_{\mathbf{post}}$ decreases, the separated signal maintains intelligibility, depending on the source signals, up to around $\Delta t_i = 50$ milliseconds.

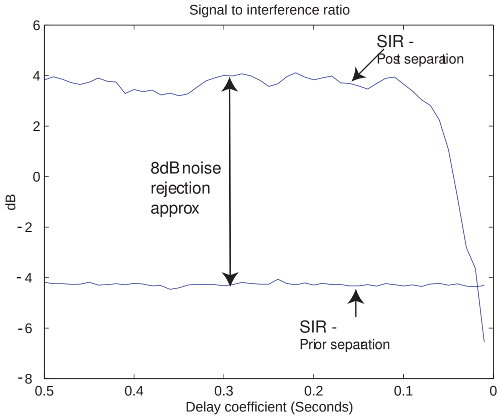


Fig. 9: The Signal to interference ratio (SIR) of the target source prior separation, SIR of the estimated source post separation over a varying delay from 0.5 to zero seconds.

VII Conclusion

This paper presented a novel approach to single channel source separation. It has been demonstrated that under simplified conditions, the theory can be applied with promising results. Further work must be undertaken to ascertain how robust the technique will be under less confined constraints, such as a greater number of sources and more realistic acoustic circumstances.

VIII Acknowledgments

Work supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project EASAIER contract IST-033902.

References

[1] P. L. Montgomery. "Modular multiplication without trial division". *Math. Computation*, 44:519–521, 1985.

[2] Barry, D., Lawlor, B., Coyle, E. "Real-time Sound Source Separation: Azimuth Discrimination and Resynthesis", AES 2004.

[3] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," *in 3rd International Conference on Independent Component Analysis and Blind Source Separation*, San Diego, CA, December 9-12 2001.

[4] Lee, D., Seung, H. "Algorithms for Non-negative Matrix Factorization", *Adv. Neural Info. Proc. Syst.* , 13, 556-562, 2001.

[5] Jourjine, A., Rickard, S., Yilmaz, O. "Blind Separation of Disjoint orthogonal signals: Demixing N sources from 2 mixtures", *ICASSP*, 2000.

[6] Hyvärinen, A., Erkki, Oja. "Independent Component Analysis: Algorithms and Applications", *Neural Networks*, 13(4-5):411-430,2000.

[7] Stautner, J.P., "Analysis and Synthesis of Music using the Auditory" Transform", *Masters Thesis*, MIT EECS Department, 1983

[8] E. Vincent, R. Gribonval, and C. Favotte,. "Performance Measurement in Blind Audio Source Separation". *IEEE Transactions on Speech and Audio Processing*, 2005.

[9] Lin, Y., Lee, D. D. and Saul, L. K. "Nonnegative deconvolution for time of arrival estimation.". *In Proceedings of the international Conference of Speech, Acoustics, and Signal Processing (ICASSP-2004)*, volume 2, pages 377-380, Montreal, Canada, 2004.