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

Mark Leddy  
*Technological University Dublin*

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

David Dorran  
*Technological University Dublin*, david.dorran@tudublin.ie

*See next page for additional authors*

Follow this and additional works at: [https://arrow.tudublin.ie/argcon](https://arrow.tudublin.ie/argcon)

Part of the Other Engineering Commons

**Recommended Citation**


This Conference Paper is brought to you for free and open access by the Audio Research Group at ARROW@TU Dublin. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, gerard.connolly@tudublin.ie, vera.kilshaw@tudublin.ie.
Authors
Mark Leddy, Dan Barry, David Dorran, and Eugene Coyle

This conference paper is available at ARROW@TU Dublin: https://arrow.tudublin.ie/argcon/4
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 of 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].

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

where $\alpha_i \geq 0$ represents the attenuation over the extra distances travelled along each reflected path
R_i, in comparison to the direct path d_i 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 ∆t_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.

R(x) (reflected path, compared to that of the direct path, due to the extra time required to traverse the reflected path, upon reaching the sensor they will appear as delayed versions of the source.

equation (2) is used.

\[ x(t) = [s_1(t) + \alpha s_1(t + \Delta t_1)] + ... + [s_2(t) + \beta s_2(t + \Delta t_2)] \]  

where \( s_i(t) \) represents sources received by the sensor at time \( t \). The value \( \Delta t_i \) 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 \( \alpha \) and \( \beta \) respectively. Hence the sensor mixture will consist of each source, and one delayed and attenuated version of each source.

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 this paper.

In theory, it should be possible to separate single sources, from a mixture of a large number of sources.
the attenuation coefficient of the delayed source is located in the center of the stereo field. Generally, same intensity in both channels, and hence will be negligible, i.e., aligned in time. If the attenuation coefficient is $\alpha_i$, its delayed version $\alpha_i(t + \Delta t_i)$ 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, i.e., $\alpha = 1$, 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 less than 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

![Fig. 4: Autocorrelation of the mixture signals. The peaks are used to estimate the delay coefficients $\Delta t_i$.](image)

![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.](image)
SIR is used. The system was tested by employing 50 linearly spaced delay values, between 0 and 500 milliseconds, 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.

$$\text{SIR}_{\text{pre}} := 20 \log_{10} \frac{|S_{\text{target}}|}{|S_{\text{mixture}} - S_{\text{target}}|}$$  \hspace{1cm} (4)$$

$$\text{SIR}_{\text{post}} := 20 \log_{10} \frac{|S_{\text{estimate}}|}{|S_{\text{estimate}} - S_{\text{target}}|}$$  \hspace{1cm} (5)$$

where $S_{\text{target}}$ is the test signal to be separated, $S_{\text{estimate}}$ represents the separated estimation of

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 500 milliseconds, 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.

$$\text{SIR}_{\text{pre}} := 20 \log_{10} \frac{|S_{\text{target}}|}{|S_{\text{mixture}} - S_{\text{target}}|}$$  \hspace{1cm} (4)$$

$$\text{SIR}_{\text{post}} := 20 \log_{10} \frac{|S_{\text{estimate}}|}{|S_{\text{estimate}} - S_{\text{target}}|}$$  \hspace{1cm} (5)$$

where $S_{\text{target}}$ is the test signal to be separated, $S_{\text{estimate}}$ represents the separated estimation of

the target signal. $S_{\text{mixture}}$ is the mixture of signals represented by equation (2).

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 $-4\text{dB}$. After applying the process described in this paper, analysis shows that an average of $+4\text{dB}$ of signal to interference ratio has been achieved, resulting in an average signal to noise ratio increase of $8\text{dB}$.

$\text{SIR}_{\text{post}}$ for a delay coefficient from 0.5 to 0.1 seconds averages about $+4\text{dB}$. For this delay timeframe, an approximate $+8\text{dB}$ noise rejection difference is maintained over $\text{SIR}_{\text{pre}}$. As the delay coefficient decreases below 0.1 seconds $\text{SIR}_{\text{post}}$ diminishes. However, even though $\text{SIR}_{\text{post}}$ 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


