# Demixing of Speech Mixtures and Enhancement of Noisy Speech Using ADRess Algorithm

**Niall Cahill, Rory Cooney, Kenneth Humphreys and Robert Lawlor**

*Department of Electronic Engineering,*
*National University of Ireland, Maynooth,*
*Maynooth,*
*Co. Kildare,*
*IRELAND*
*E-mail: niall.cahill@eeng.nuim.ie*

**This paper describes the ability of the Azimuth Discrimination and Resynthesis algorithm (ADRess) to separate multiple speech signals from two mixtures in a simulation environment. ADRess exploits the spatial signature of each of the contributing speech sources to demix the mixtures. Speech sentences taken from the TIMIT database and noise signals from the NOISEX database were mixed synthetically to create pairs of mixtures. ADRess can exploit the spatial signature of noise and speech sources to remove or isolate them from a mixture. To simulate the spatial location of different sources the relative attenuation and phase difference of each source between the two mixtures were manipulated. This was performed for numerous different angles of arrival so as to robustly test the algorithm. Objective measures and promising informal listening test results show the suitability of ADRess for cleaning noisy speech mixtures and document the performance of ADRess for speech mixtures with different numbers of sources.**

**Keywords – Sound Source Separation, Speech Enhancement.**

## I INTRODUCTION

Sound source separation algorithms attempt to separate sound mixtures that contain a plurality of different sound sources into the constituent sources. This problem is sometimes referred to as the "cocktail party problem".

The "cocktail party effect" refers to the innate ability of humans to discern individual sources of sound despite being in the presence of a multitude of interfering sound sources. Auditory scene analysis (ASA) [1] is the term used to describe how humans are capable of segregating different sound sources, which may overlap with each other in both the time and frequency domain. To describe how humans achieve this Bergman [1] identified two forms of organisation performed by humans, simultaneous and sequential organisation. Simultaneous organisation deals with the separation and grouping of sounds occurring at the same time, this corresponds to grouping in the frequency domain. Sequential organisation deals with sounds that occur. at different times. These two mechanisms enable humans to group auditory events according to the common fate principle. This principle states that a set of auditory cues occurring simultaneously or sequentially in the frequency or time domain may be grouped as a single source. Computational Auditory scene analysis (CASA) attempts to use this model to design algorithms capable of sound source separation. However modelling the processing of the brain is a major obstacle to practical algorithms.

A statistical approach to this problem is that of Independent component analysis (ICA) [2]. This can be formally defined as follows: Consider a number of sources $s_i(t)$, which are linearly mixed using mixing matrix $A$ with coefficients $a_{ij}$ producing mixtures $x_i(t)$. The mixing equation can be written as,

$$x = As. \tag{1}$$

The aim of ICA is to find a separation matrix $W$ that is the inverse of the mixing matrix $A$.

$$u = Wx = WAs = A^{-1}As = s. \tag{2}$$

Where $u$ contains the separated signals. The limitation of ICA is that the number of sources must equal the number of sensors to be able to calculate $A^{-1}$ in equation 2. This is impractical in certain environments where there may be no prior knowledge of the number of sources.

Separating mixed speech signals when there are more speech sources than mixtures can be achieved using the Degenerate Unmixing Estimation Technique (DUET)[3]. This technique uses only two mixtures of the speech to separate. DUET is based on the assumption that speech signals are *sparse* in the time-frequency domain, and appear disjoint in this representation. This property of speech is known as approximate W-disjoint orthogonality (WDO) [4]. The task of separation then reduces to deciding which time-frequency points belong to which source. DUET uses a weighted histogram of relative attenuation and phase-difference between two mixtures in the time-frequency domain, to associate a relative attenuation and phase difference pair to each source. A distance metric is then used to decide which source each time–frequency point belongs to. It then applies a binary mask to these time frequency points zeroing all points not deemed part of the same source.

The Azimuth Discrimination and Resynthesis algorithm (ADRess) [5] uses only two mixtures to demix numerous music sources. ADRess has been developed to demix stereo music recordings [5]. ADRess has the capability to segregate time frequency points based on relative channel differences therefore it is capable of separating speech mixtures from two microphone mixtures. In this paper the ADRess algorithm is presented in a configuration appropriate for separating speech and speech mixed with noise in an anechoic environment. The issues involved in the conversion from stereo music recording to two-microphone speech recording are presented. In section III a Modified ADRess algorithm is introduced that is adapted to the two-microphone speech mixture setting. The performance of Modified ADRess for both tasks is evaluated using both subjective and objective metrics. Applications of this work would include enhancement of mobile phone speech or degenerate sound source separation.

## II ADRess Background

The Azimuth discrimination and resynthesis algorithm (ADRess) was developed to separate stereo musical recordings into independent constituent sources that comprise the mixtures. ADRess utilizes the process of synthesising the sensation of space in stereo recordings. This process usually involves recording each instrument separately in a recording studio and then using a panoramic potentiometer (pan pot), an interaural intensity difference (IID) is created between the left and right channel of the stereo signal. The pan pot simply increases the presence of one source in one channel relative to the other by scaling the source appropriately. The scaled sources for each channel are then summed together to create the left and right channel signals. Listeners perceive IID as the apparent location of the sources along a horizontal stereo field from left to right.

## III ADRess Methodology

ADRess assumes the following discrete time mixing model,

$$l(n) = \sum_{i=0}^{j-1} al_i s_i(n) \quad , n = 0,..,N\text{-}1, \tag{3}$$

$$r(n) = \sum_{i=0}^{j-1} ar_i s_i(n) \quad , n = 0,..,N\text{-}1. \tag{4}$$

Where $l(n)$ and $r(n)$ are the left and right channel stereo signals , $al_i$ and $ar_i$ are the left and right panning coefficients, $s_i(n)$ is the $i^{th}$ independent source, $N$ is the length of the channels in samples and $j$ is the number of sources. The algorithm takes these two signals as its initial input data and then divides them into short overlapping frames. These frames are transformed into the frequency domain using the Fourier Transform yielding the following expressions.

$$l_f(k) = \sum_{n=0}^{N-1} w(n) l(n) e^{\frac{-j2\pi kn}{N}}, \tag{5}$$

$$r_f(k) = \sum_{n=0}^{N-1} w(n) r(n) e^{\frac{-j2\pi kn}{N}}. \tag{6}$$

Where $k$ is the frequency point, $w$ is a windowing function, usually a Hamming window and $N$ is now the frame size.

From equations (3) and (4) the ratio of the left and right panning coefficients $al$ and $ar$ for the $i^{th}$ source can be expressed as,

$$g(i) = al_i / ar_i . \tag{7}$$

Similarly,

$$al_i = g(i).ar_i, \tag{8}$$

and

$$al_i - g(i).ar_i = 0, \tag{9}$$

where $g(i)$ is the intensity ratio. The relationship in equation (7) implies that scaling $r_f$ by $g(i)$ and subtracting this value from $l_f$ results in the source $i$ being cancelled from $l_f$. If this is performed for a

range of different intensity ratios other sources present in each channel for a particular $g$ will cancelled.

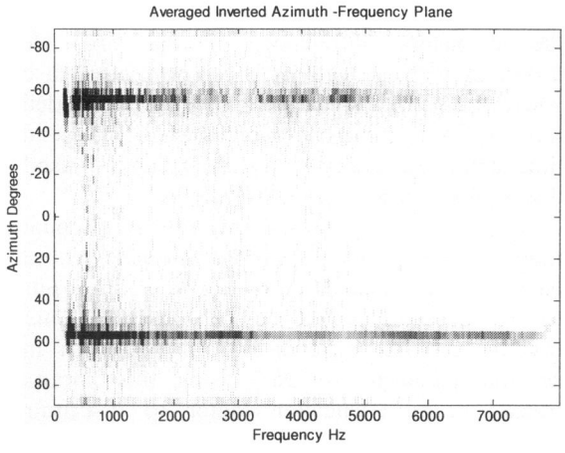


**Figure 1:** Inverted Frequency Azimuth Plane of a speech mixture averaged over 250 frames. The mixture contains two sources of different azimuth.

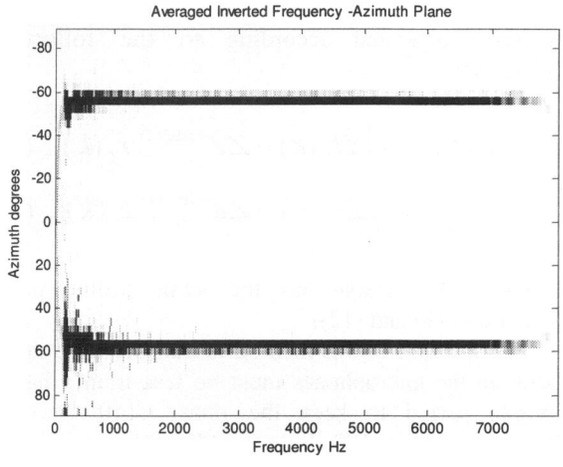


**Figure 2:** Two superimposed Averaged Inverse Frequency Azimuth Planes of two independent speech sources with different azimuths.

Similarly $r - g(i).l$, will scale the left source which when subtracted from the right source will remove sources, which predominate on the left channel, from the right. When the panning coefficients are unknown as is the case of a stereo recording a set of gain scale factors must be defined. The gain scale factors are defined as follows,

$$g(i) = i.(1/\beta), \tag{10}$$

for all $i$ and for $0 \leq i \leq \beta$ where $i$ and $\beta$ are integer values.

Right and left channel azimuth-frequency planes are created according to the following equations.

$$Azr(k,i) = \left| l_f(k) - g(i).r_f(k) \right|, \tag{11}$$

$$Azl(k,i) = \left| r_f(k) - g(i).l_f(k) \right|, \tag{12}$$

for all $i$ and $k$ where $0 \leq i \leq \beta$, and $1 \leq k \leq N$. This will result in a $N$ by $\beta$ matrix. Combining $Azr$ and $Azl$ creates the azimuth frequency plane of the mixture. For each frequency along the azimuth axis, there exist peaks of varying magnitude, resulting from the gain scale subtraction process. For each frequency, these peaks converge to a minimum value, which corresponds to the location of that frequency within the azimuth plane provided that the signals are orthogonal (that is, no other source produced this frequency at this exact moment in time). This is true for each frequency component in the signal, with each component converging to the same point in the azimuth plane. This can be seen in the inverted azimuth plane of Figure 2. For the purpose of re-synthesis, the convergent values in the each azimuth plane are inverted. A peak is assigned to the location of the null (or minimum value) having a magnitude equal to the difference between the value of the null and the maximum value of the azimuth plane at that frequency. All other points in the azimuth plane are zeroed. This is performed on a frame-by-frame basis.

$$Azr(k,i) = \begin{cases} Azr(k)_{\max} - Azr(k)_{\min}, & if\ Azr(k,i) = Azr(k)_{\min} \\ 0, \ Otherwise. \end{cases} \tag{13}$$

$$Azl(k,i) = \begin{cases} Azl(k)_{\max} - Azl(k)_{\min}, & if\ Azl(k,i) = Azl(k)_{\min} \\ 0, \ Otherwise. \end{cases} \tag{14}$$

To resynthesis a portion of this *Azl–Azr* plane an azimuth point is chosen. If an azimuth is chosen where a source has been panned numerous magnitude peaks will be situated along the frequency axis at this azimuth corresponding to frequencies where this source contains energy. These peaks are then used with the original bin phases to synthesise the source present at that azimuth.

In practice each source in a mixture is not strictly orthogonal with every other source. This complication leads to certain frequencies containing energy from multiple sources. The peaks of these frequencies drift away from a source position and locate at an erroneous azimuth where there may or may not be a source. This is illustrated in Figure 1 where the inverted frequency azimuth plane of a two-source speech mixture is averaged over each frame in the mixture. The resultant matrix plotted in Figure 1 showing the spread of time-frequency points for two sources that are not completely orthogonal. This can be contrasted with Figure 2 where the same two sources are analysed separately, yielding two distinct frequency azimuth planes. Superimposing the two frequency azimuth planes on each other will show the position of time frequency

points of the two sources if they were completely orthogonal.

The *azimuth-smearing* phenomenon in Figure 1 results in frequencies being excluded from the resynthesis of the target source. To include these frequencies, which contain energy other than the energy of the target source, an "azimuth subspace width" $H$ is defined. This permits including peaks that have drifted away from the target azimuth in the resynthesis of the source. An extra term the "discrimination index" $d$ is also defined at this stage. Collectively $H$ and $d$ will define what portion of the azimuth frequency plane will be used for resynthesis.

$$Yr(k) = \sum_{i=d-H/2}^{i=d+H/2} Azr(k,i), \quad (15)$$

$$Yl(k) = \sum_{i=d-H/2}^{i=d+H/2} Azl(k,i) \quad (16)$$

The phase and magnitude component of each bin are combined and converted from polar to rectangular form. The Inverse Fourier Transform is then applied to transform these points from the frequency to the time domain.

$$X(n) = \frac{1}{N}\sum_{k=1}^{N} X(k) e^{\frac{-j2\pi kn}{N}}, \quad (17)$$

This is performed for each frame with successive frames recombined using a simple overlap and add scheme.

## IV SPEECH MIXTURES

Speech signals recorded in an anechoic environment with two microphones have different properties to stereo music recordings. Stereo music recordings use an intensity difference between the left and right channel to position sources to different locations. However in a speech mixture-recording scenario (conference room) sources are located in different positions around a room, the speech sources are received at two separated microphones with intensity and time delay differences assuming the path length for a source to two separated microphones are different lengths. For one source and two microphones this can be expressed mathematically as follows,

$$mic_1(t) = s(t), \quad (18)$$

$$mic_2(t) = a_1 s(t) e^{-j\omega d}, \quad (19)$$

where $mic_1$ and $mic_2$ are two separated microphone signals, $s$ is the source, $a$ is the attenuation factor between the two microphones, $d$ is the delay difference between the two microphones and $\omega$ is the complex frequency vector from dc to the sampling frequency. Assuming the source arrives at microphone one before microphone two, the ratio of the attenuation difference, or the time delay difference between the signals can be used to discern where the speech signal originated. If extended to mixtures with numerous sources each independent time delay and attenuation factor can be used to discern the signals.

ADRess already uses a gain factor/attenuation ratio to expose where sources in a music recording have been positioned. Applying this approach to speech mixtures would be sub-optimal. This is because the time difference of arrival is a more accurate parameter to use. Using this parameter instead of an attenuation factor to discriminate between sources would result in a more accurate spatially based separation.

To separate speech mixtures ADRess has been modified to utilize the relative time delay differences between sources impinging on two microphones. The frequency azimuth plane of equations (11) and (12) is now generated according to the following equations,

$$Azr(k,i) = \left| \angle l_f(k) - \angle e^{-j\omega g(i)} . r_f(k) \right|, \quad (20)$$

$$Azl(k,i) = \left| \angle r_f(k) - \angle e^{-j\omega g(i)} . l_f(k) \right|, \quad (21)$$

where each variable has the same value as in equations (11) and (12).

A constraint of this method is that the time delay between the microphones must be less than a half a sample period to keep the phase shift at each frequency below 180 degrees. Maintaining a small distance between microphones during recording obviates this. This also means that the bandwidth of the speech is related to the maximum allowed distance between the microphones. For a mobile phone application the bandwidth is not likely to exceed 4 Khz (8 Khz sample rate) and so a microphone separation of up to about 8.3 cms could be used.

## V EXPERIMENTS

### a) *Experimental Set-up*

To show that the Modified ADRess (M-ADRess) can be used to separate speech mixtures we present results in this section for synthetically mixed and panned signals. All data generated and all processing was performed using MATLAB™. To test M-ADRess mixtures of various orders were created, four two-source mixtures, one three-source mixture and a five-source mixture. Furthermore, different types of mixtures were also generated, mixtures with speech only and mixtures with speech and coloured

| Noise and Speech Positions | Input SNR (dB) | Output SNR (dB) | SNR gain | MOS rating |
|---|---|---|---|---|
| Speech 45 – Factory Noise 180 | -12.3701 | 5.7818 | 18.1519 | 1 |
| Speech 0 – Factory Noise2 0 | -14.6881 | 7.900 | 22.5881 | 1.6 |
| Speech 180 – Speech Babble 45 | -8.924 | -2.9720 | 5.952 | 1.5 |
| Speech 45 – Volvo Car noise | -15.3738 | 9.1693 | 24.5431 | 1.6 |

**Table 1:** Table of Subjective and Objective Results for speech separated from noise using M-ADRess

noise. The noise samples used were taken from the NOISEX speech database and the speech sentences from the TIMIT database. Five angles of arrival were chosen -90°, -45°, 0°, 45° and 90°. All mixtures contained sources placed at one of these angles on a horizontal plane 1 m equidistant from both microphones. The microphones were placed 2 cm from each other. The ideal attenuation and phase characteristic for sources placed at these angles relative to each microphone were used to simulate the mixtures. Speech sentences taken from the TIMIT speech database were assigned to angles, the time delay and attenuation factor for each angle was applied to each source relative to each microphone. Each time delay and attenuation factor is dependent on the angle the speech sentence was paired with and the distance from the microphone. The scaled and delayed version of each source for both microphones are then added to create the left and right microphone mixtures. These mixtures were then passed through the ADRess algorithm where the resultant azimuth-frequency plane was scanned for the best possible rendition of the target sources. The frame length used was 1024 samples, the sampling rate was 16 kHz, the analysis step size 128 samples and azimuth resolution $\beta$ was set to 20.

### b) *Subjective Evaluation of Processed Speech*

To assess the subjective quality of the processed speech the Mean Opinion Scoring (MOS) technique was employed. The MOS method is a common technique used to quantify the perceived quality of processed speech by a panel of subjects. Obtaining MOS ratings generally involves selecting a group of subjects and instructing each subject to listen to a sample of the processed speech and then choose a rating that best fits their opinion of the speech .The rating preferences are then collected and averaged over the number of listeners in the panel. For this analysis a panel of 10 subjects was used

The subjective test protocol for the mixtures of speech only and the subjective test for speech and noise mixtures were slightly different. For speech mixtures the best rendition of the original speech was extracted and presented to each subject. The subject was then asked to evaluate the processed speech according to the ratings in Table 2.

| MOS rating | Description |
|---|---|
| 1 | Perfect |
| 2 | Minor artifacts or interference |
| 3 | Distorted but intelligible |
| 4 | Very distorted and barely intelligible |
| 5 | Not intelligible |

**Table 2:** MOS rating descriptions for Speech Quality Listening Test.

The rating descriptions in Table 2 were chosen to evaluate the effect of processing artefacts (musical noise, overlapping time-frequency points etc) on the perceived quality of the output speech. Such a test also indicates the effect of increasing mixture order on the subjective quality of the separated speech. The MOS Ratings are displayed in Table 4.

To evaluate the performance of the algorithm for mixtures of speech mixed with coloured noise an alternative test protocol was used. Each subject listened to the noisy speech signal followed by the processed speech signal, enabling the listener to compare the input and output speech, producing a subjective judgement on the improvement. The following rating descriptions were used.

| MOS rating | Description |
|---|---|
| 1 | Much improved |
| 2 | Improved |
| 3 | Different but unimproved |
| 4 | Slightly worse |
| 5 | Worse |

**Table 3:** MOS rating descriptions for Speech Enhancement Listening Test.

The results of the MOS test are shown in Table 1

### c) *Objective Evaluation of Processed Speech*

To objectively assess the denoised speech two performance indicators were used, an input signal to noise ratio and an output signal to noise ratio. The input SNR in equation (22) determines the ratio of intensities between a source and all other noise sources prior to processing. The output SNR ratio in equation (23) measures the efficiency of the algorithm in removing the other sources from the target signal. Each measure was evaluated on an overlapping frame-by-frame basis.

$$SNR_{input} = 10\log\left[\frac{\sum_{n=1}^{K} s^2(n)}{\sum_{n=1}^{K} d^2(n)}\right] \qquad (22)$$

$$SNR_{output} = 10\log\left[\frac{\sum_{n=1}^{K} s^2(n)}{\sum_{n=1}^{K}[s(n)-s'(n)]^2}\right] \quad (23)$$

where $K$ is the number of frames, $s(n)$ is the original source signal, $s'(n)$ is the rendition of the original source signal from ADRess algorithm and $d$ is the sum of all interfering sources . These two measures were subsequently combined to quantify an SNR gain for the algorithm for each speech sentence. These measures are tabulated in Table 1 and Table 4 for several mixture orders.

$$SNR_{gain} = \frac{SNR_{output}}{SNR_{input}} \quad (24)$$

## VI DISCUSSION

From Table 4 it is generally observed that the M-ADRess algorithm can indeed separate mixtures with two to three sources. This is shown in the MOS ratings, which show an acceptable level of distortion for mixtures of this order. Mixtures with more than three sources show an unacceptable decline in speech quality. The results shown in Table 1 suggest that the M-ADRess algorithm is suited to speech enhancement. This is reflected in the MOS values, which indicate the perceptual improvement in the processed speech using M-ADRess.

The MOS ratings from Table 4 indicate that as the source order increases there is a subjective decline in quality. This is to be expected as it has been shown that as the order of the mixture increases, the W-disjoint orthogonality of the speech signals in the mixture decrease [3]. This implies that there will be increased overlap of time-frequency points leading to increased Azimuth smearing.

All the above experiments were performed in a simulated echoic environment. This assumption is used to simplify the task of separating the mixtures. In an echoic environment there will be reflections from walls obstacles etc; these reflections will create multiple paths to the microphones making the separation task more complicated. A multi-path signal will have a different spatial signature for each path, whereas in an anechoic environment there is only a direct path from each source to both microphones. The energy of a source in an anechoic environment will thus be focused about an individual azimuth, while the energy of an echoic signal will appear spread across the azimuth frequency plane. Efficient resynthesis is achieved using the ADRess algorithm in an anechoic environment however resynthesis in real environments requires further work.

## VII CONCLUSIONS

We have demonstrated that by configuring the ADRess algorithm to discriminate based on time delays only anechoic speech mixtures can be successfully separated. It has been shown that this modified ADRess algorithm has the capability to separate speech mixtures using only two spatially independent examples of the mixtures. Also highlighted in this paper is the suitability of the modified ADRess algorithm for speech enhancement.

| Source positions | MOS rating |
|---|---|
| 0 | 2.2 |
| 90 | 3.2 |
| 0 | 2.8 |
| 45 | 2.5 |
| 0 | 2.3 |
| -45 | 3.5 |
| 0 | 2.5 |
| -90 | 2 |
| 90 | 2.7 |
| 0 | 2.7 |
| -90 | 3.7 |
| -90 | 4.6 |
| -45 | 4.9 |
| 0 | 3.3 |
| 45 | 4.7 |
| 90 | 3.4 |

**Table 4:** Subjective results for speech separated from mixtures of various orders.

## VIII REFERENCES


[1] Bregman, A. S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound. *Cambridge, MA. The MIT Press.*

[2] A. Hyvarinen, "Survey on Independent Component Analysis", *Neural Computing Surveys 2, 94-128.*

[3] A. Jourjine, S. Rickard, O. Yilmaz, "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures". *IEEE International Conference on Acoustics and Speech and Signal Processing*, June2000

[4] Rickard, S.; Yilmaz, Z.;" On the approximate W- disjoint orthogonality of speech" *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02).* IEEE International Conference on.Volume 1, 2002

[5] D Barry, B Lawlor, E Coyle, "Sound Source Separation ": Azimuth Discrimination and Resynthesis", *Proc of the 7th int. conference on Digital Audio Effects (DAFX-04)*, Naples, Italy, October 5-8, 2004.