2022-06-10

Soft-Mask De-Mixing for Anechoic Mixtures

Swarnadeep Bagchi
Technological University Dublin, d18128352@mytudublin.ie

Ruairí de Fréin
Technological University Dublin, ruairi.defrein@tudublin.ie

Recommended Citation

S. Bagchi and R. De Fréin, "Soft-Mask De-Mixing for Anechoic Mixtures," 2022 33rd Irish Signals and Systems Conference (ISSC), Cork, Ireland, 2022, pp. 1-6, doi: 10.1109/ISSC55427.2022.9826179.

Funder: SFI

Soft-Mask De-Mixing for Anechoic Mixtures

S. Bagchi and Ruairí de Fréin
Ollscoil Teicneolaíochta Baile Átha Cliath,
Campas na Cathrach,
Ireland.
web: https://robustandscalable.wordpress.com

in: 2022 33rd Irish Signals and Systems Conference (ISSC). See also BibTEX entry below.

BibTEX:

```
@article{deFrein22Soft,
  author={Bagchi, Swarnadeep and de Fr\'{e}in, Ruair\'{i}},
  journal={2022 33rd Irish Signals and Systems Conference (ISSC)},
  title={Soft-Mask De-Mixing for Anechoic Mixtures},
  year={2022},
  volume={},
  number={},
  pages={1-6},
  doi={10.1109/ISSC55427.2022.9826179},
  url = {https://ieeexplore.ieee.org/abstract/document/9826179}
  }
```

Soft-Mask De-Mixing for Anechoic Mixtures

Swarnadeep Bagchi
Technological University Dublin, Ireland
D18128352@mytudublin.ie

Ruairí de Fréin
Technological University Dublin, Ireland
ruairi.defrein@tudublin.ie

***Abstract*—This paper extends a computationally efficient, soft-mask based source separation (SS) technique called Redress, to anechoic mixing scenarios. SS methods are an integral part of hearing aid research. We call the resulting method D-Redress. In its original form, Redress was intended for instantaneous mixing scenarios. Numerical evaluations demonstrate that soft-mask based techniques reduce the level of artifacts in the separated speech. Monte Carlo trials on 1000 real speech mixtures demonstrate that the D-Redress successfully extends Redress in terms of Overall-Perceptual (OPS), Target-Perceptual (TPS) scores and Human-Ear Intelligibility (HEI).**

***Keywords*—Source-Separation, anechoic, relative attenuation, Overall-Perceptual Score (OPS), Target-Perceptual Score (TPS), Human-Ear Intelligibility (HEI)**

I. Introduction

A discrete-time, stereo anechoic de-mixing scenario consisting of J sources is given as:

$$x_1[n] = \sum_{j=1}^{J} s_j[n] \tag{1}$$

$$x_2[n] = \sum_{j=1}^{J} \alpha_j s_j[n - \delta_j] \tag{2}$$

Here n is the discrete time index, $1 \leq n \leq N$ and $n \in \mathbb{Z}_+$. The total number of discrete time is N. Source $s_j[n]$ in mixture $x_2[n]$ is relatively attenuated by α_j, where $0 < \alpha_j \leq 1$ and delayed by δ_j samples, relative to $x_1[n]$. Sources are pan-mixed, this gives the sources a location in the stereo-field in line with their corresponding attenuation coefficients a_j. Typical well known techniques for separating out the sources from $x_1[n]$ and $x_2[n]$ are ESPRIT [1], DUET [2], TIFROM [3] and the DEMIX [4] algorithms. DESPRIT [5] is an extension of ESPRIT. A family of power-weighted estimators was introduced in [6]. It demonstrates that these estimators can be unified into one statistical framework. These approaches were built on initial contributions in the area of Independent Component Analysis (ICA) [7]. The essence of ICA based methods lies in the exploitation of the sparsity of the sources in some transform domain [8]. The sources are sparse implies that they are in some sense already separated. SS can then be achieved via a binary or hard mask as shown in Eqn. 3. Hard-masking techniques assume that a Time-Frequency (TF) bin belongs to any one source. According to the authors in [9], if all the TF bins that correspond to a particular source are determined, then the sources can be separated.

$$\mathbf{M}[k,\tau] = \begin{cases} 1, & \text{if TF bin is in the source} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Here, the mask is $\mathbf{M}[k,\tau]$ and k and τ are the discrete frequency and time indices of the TF representation, respectively. We contribute to an extension of the Redress algorithm [10] to anechoic mixtures, which does not significantly increase the computational cost of it. This approach is called Delayed-Redress (D-Redress). Redress is a computationally efficient soft-masking technique for source separation. The sources are relatively attenuated on different channels. The soft-mask based methods generate a ratio-mask, its value lies anywhere in between $0 \leq \mathbf{M}[k,\tau] \leq 1$. The Redress technique localizes the inter-aural Intensity Difference (IID) cue of the j^{th} source at an attenuation location in a frequency-attenuation matrix. This location equals the attenuation coefficient α_j in the stereo-field. The underlying idea of soft-mask techniques is to allocate appropriate spectral-power to the sources in their corresponding TF-bins. This allocation of spectral-power is in line with the source's attenuation coefficient α_j. The hard-mask counterpart of Redress is AdRess [11]. The downside of these methods are that they are restricted to the instantaneous mixing scenario. An extension of the AdRess to anechoic scenario was recently contributed in [12]. This is known as D-AdRess. In this paper, we have compared our proposed D-Redress with Redress, AdRess and D-AdRess. We hypothesize that soft-mask based methods decrease the level of artifacts introduced in the separated speech. Typically, interference of other sources increase as we move from hard-masking to soft-masking techniques.

This paper is organized as follows. In Section II the effect of phase on delayed sources is explained. In Section III, Redress algorithms is detailed. D-Redress is defined in Section IV. In Section V, we evaluate the performance of D-Redress by considering the task of separating real speech signals, where up to four sources are present in the mixture. We have compared D-Redress algorithm with other known techniques. The paper finishes with the concluding remarks in Section VI.

II. Phase-Aware Intuition

SS techniques take the first step of computing the TF representation of the mixtures. Typically, the synchronized short-time-Fourier-transform (sSTFT) [13], Wavelet Transform

978-1-6654-5227-4/22/$31.00 ©2022 IEEE

and Wigner-Ville distribution are used. The TF transform of an arbitrary discrete-time speech mixture $x[n]$, where $1 \leq n \leq N$, is $\mathbf{X} : x[n] \longmapsto \mathbf{X}[k,\tau] \in \mathbb{C}$. Here, $1 \leq k \leq K$ and $1 \leq \tau \leq T$. The total number of discrete frequencies and time frames are K and T, respectively. The magnitude and the phase spectrum of $x[n]$ is given as $|\mathbf{X}[k,\tau]|$ and $\angle\mathbf{X}[k,\tau]$, respectively. Now, $\mathbf{X}[k,\tau]$ of the matrix $\mathbf{X} \in \mathbb{C}^{K\times T}$ can be written in a phasor form as $\mathbf{X}[k,\tau] = |\mathbf{X}[k,\tau]|e^{j\theta}$, where θ is the phase of the TF bin. Again, a temporal signal incurring a relative delay of δ samples, $x[n-\delta]$, is represented in the TF-domain as $\mathbf{X}[k,\tau] = |\mathbf{X}[k,\tau]|e^{j\theta}e^{-j\Omega_k\delta} = |\mathbf{X}[k,\tau]|e^{j(\theta-\Omega_k\delta)}$. Here, the discrete angular frequency (in rads/sample) is $\Omega_k = \frac{2\pi}{K}k$ for a K-point DFT and the k^{th} index. The phase of $\mathbf{X}[k,\tau]$ becomes $e^{j(\theta-\Omega_k\delta)}$. Both the magnitude $|\mathbf{X}[k,\tau]|$ and the phase $\angle\mathbf{X}[k,\tau]$ spectra are used for speech analysis and enhancement [14]. We state that signal separation becomes difficult when the mixture $\mathbf{X}[k,\tau]$ is influenced by a phase quantity of $e^{j(\theta-\Omega_k\delta)}$ than when influenced by only $e^{j\theta}$. Therefore, source-separation from a mixture with the help of $|\mathbf{X}[k,\tau]|$ alone won't be possible in an anechoic mixing scenario. In the Section IV, we shall see how delays influence the traditional analysis of the TF representation for Redress algorithm.

III. Redress Algorithm

Let us consider two synthetic signals to motivate our approach. They consist of sinusoids $s_1[n]$ and $s_2[n]$. Source $s_1[n]$ is made up of frequencies $f_1 = 100$ Hz and $f_3 = 300$ Hz; $s_1[n] = \sin(2\pi f_1 t) + \sin(2\pi f_3 t)$. Source $s_2[n]$ is made up of frequencies $f_2 = 200$ Hz and $f_4 = 400$ Hz; $s_2[n] = \sin(2\pi f_2 t) + \sin(2\pi f_4 t)$. They compose two instantaneous mixtures $x_1[n]$ and $x_2[n]$. Their TF representations are $\mathbf{X}_1[k,\tau]$ and $\mathbf{X}_2[k,\tau]$, respectively. In the mixing matrix, the attenuation coefficients are $\alpha_1 = 0.2$ and $\alpha_2 = 0.8$. Sources are non-overlapped in frequencies. Considering an instantaneous scenario, Redress takes the magnitude spectrum of one mixture $|\mathbf{X_1}|$ and scales it relative to $|\mathbf{X_2}|$ and vice-versa, such that the difference of the magnitude of the TF-representation is zero or null.

$$\begin{aligned} g \leftarrow \text{find}\, \big(\mathbf{A_1} &= \big|\mathbf{X}_1[k,\tau] - g\mathbf{X}_2[k,\tau]\big| \approx 0, \\ \text{and } \mathbf{A}_2 &= \big|\mathbf{X}_2[k,\tau] - g\mathbf{X}_1[k,\tau]\big| \approx 0\big)\,. \end{aligned} \tag{4}$$

Redress constructs a frequency-attenuation matrix, $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2] \in \mathbb{R}^{K\times M}$, depicted in Fig. 1a. Here, K and M are the number discrete frequencies and size of the attenuation range, respectively. The attenuation parameter is g where $0 < g \leq 1$. The set of attenuation values examined is:

$$g = \{g_1, g_2, \ldots\ldots, 1\} \tag{5}$$

We have considered the size of the set g to be $G = \frac{M}{2} = 100$, for low computational complexity. The Redress technique formulates the matrix $\mathbf{A}$ in a way to decompose it into its TF-spectra, $\mathbf{W}$, and attenuation component $\mathbf{H}$, so that $\mathbf{A} \approx \mathbf{WH}$. The author of the paper [15] applied a Non-Negative Matrix Factorization (NMF) technique to decompose $\mathbf{A}$, where $\mathbf{A} \in {\mathbb{R}_+}^{K\times M}$, $\mathbf{W} \in {\mathbb{R}_+}^{K\times L}$, $\mathbf{H} \in {\mathbb{R}_+}^{L\times M}$ and $L \ll M$, $L \ll K$. The approach struggled to separate the sources if the number of sources were more than the mixtures. In the subsequent paper [10], the same author formulated the problem as a non-negative Quadratic Program given as: $\min\limits_{\mathbf{W}} ||\mathbf{A} - \mathbf{WH}||_F^2$, keeping $\mathbf{H}$ fixed, $\mathbf{W}$ is updated using a multiplicative-update optimization method, as proposed in [16]. For simplicity, we consider the two mixtures $x_1[n]$ and $x_2[n]$ having discrete Fourier transforms $\mathbf{X}_1[k]$ and $\mathbf{X}_2[k]$, respectively. Then the Left-Hand Side (LHS) matrix, $\mathbf{A_1}$, of $\mathbf{A}$, is formulated as:

$$\mathbf{A_1} = \big|\mathbf{X}_1[k] - g\mathbf{X}_2[k]\big| \tag{6}$$

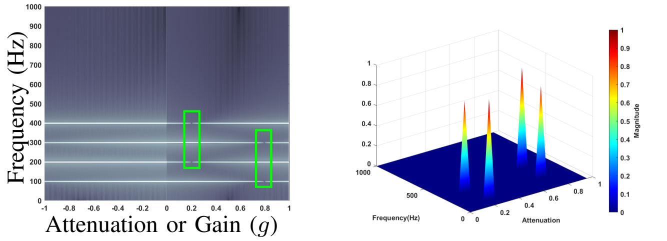
We consider only the positive frequencies due to the symmetry of the TF representation. Solving Eqn. 6 gives us the below:

$$\begin{aligned} \mathbf{A_1} = \Bigg|&\Bigg(\delta[k-f_1] + \delta[k-f_3]\Bigg)\Bigg(1 - g\alpha_1\Bigg) \\ &+ \Bigg(\delta[k-f_2] + \delta[k-f_4]\Bigg)\Bigg(1 - g\alpha_2\Bigg)\Bigg| \end{aligned} \tag{7}$$

In a similar way, the Right-Hand Side (RHS) of $\mathbf{A}$ is given as:

$$\begin{aligned} \mathbf{A_2} = \Bigg|&\Bigg(\delta[k-f_1] + \delta[k-f_3]\Bigg)\Bigg(\alpha_1 - g\Bigg) \\ &+ \Bigg(\delta[k-f_2] + \delta[k-f_4]\Bigg)\Bigg(\alpha_2 - g\Bigg)\Bigg| \end{aligned} \tag{8}$$

The derivation of Eqn. 7 and Eqn. 8 is given in paper [10]. If $g = \frac{1}{\alpha_1}$ or $g = \frac{1}{\alpha_2}$, and $g = \alpha_1$ or $g = \alpha_2$, respectively, then null points are observed at frequencies $k = f_1 = 100$ Hz, $k = f_2 = 200$ Hz, $k = f_3 = 300$ Hz, $k = f_4 = 400$ Hz in the matrix $\mathbf{A}$, provided $g \leq 1$. This is depicted in Fig. 1a. Our anechoic stereo-mixture is composed of one direct path and the other with relative attenuations and delays, nulls appear on the RHS $\mathbf{A_2}$, as depicted in Fig. 1b. These nulls determine the exact amount of spectral power by which a source gets cancelled. Peaks on these nulls give an estimate of the power content of the constituent sources. The attenuation estimates are derived where the power of the sources gets cancelled-out. We observe that this is in line with the relative attenuation of the sources present in the mixture. In Fig. 1b, the sources are located at: $g = \alpha_1 = 0.2$ for $s_1[n]$ and $g = \alpha_2 = 0.8$ for $s_2[n]$. The efficacy of Redress also lies in SS for overlapping frequencies. Real-world speech utterances are overlapped in frequencies. Let the higher frequencies of the two sources overlap at $f_3 = f_4 = 300$ Hz. Redress assigns a TF bin to multiple sources. Spectral power is reallocated to the same bin based upon the attenuation positions α_j. Once the α_j's are estimated, Redress then pre-computes the attenuation



(a) (b)

Fig. 1. Here, figure (a) illustrates the frequency-attenuation matrix $\mathbf{A}(k, \alpha)$ where the nulls appear in the RHS, $\mathbf{A}_2$, for anechoic scenarios. Figure (b) depicts the RHS $\mathbf{A}_2$ of the frequency-attenuation matrix. Nulls appears at attenuation coefficients $\alpha_1 = 0.2$ and $\alpha_2 = 0.8$. The position of these nulls are the attenuation estimates, α_j, of the sources present in the mixture.

component $\mathbf{H} = [\mathbf{H}[1,:], \mathbf{H}[2,:]]^T$. We place the first source $s_1[n]$ in the first row of $\mathbf{H}$

$$\mathbf{H}[1,:] = [|1 - g_1\alpha_1|, |1 - g_2\alpha_1|, \ldots, |1 - g_{\frac{M}{2}}\alpha_1|, |\alpha_1 - g_1|, |\alpha_1 - g_2|, \ldots, |\alpha_1 - g_{\frac{M}{2}}|] \quad (9)$$

and the second source $s_2[n]$ is placed in the second row

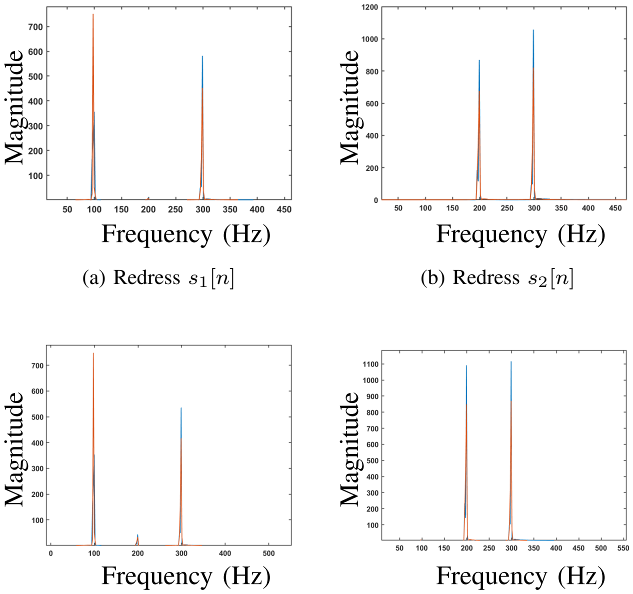
$$\mathbf{H}[2,:] = [|1 - g_1\alpha_2|, |1 - g_2\alpha_2|, \ldots, |1 - g_{\frac{M}{2}}\alpha_2|, |\alpha_2 - g_1|, |\alpha_2 - g_2|, \ldots, |\alpha_2 - g_{\frac{M}{2}}|] \quad (10)$$

For a two-source case, we have $\mathbf{H} \in \mathbb{R}^{2 \times M}$. The TF-spectra are captured by the matrix $\mathbf{W}$. Each column of $\mathbf{W}$ corresponds to the TF-spectrum of a source. For L sources, $\mathbf{W}$ shall have L columns. Fig. 2a depicts the estimated magnitude-spectrum of source $s_1[n]$, energy is assigned at frequencies $f_1 = 100$ Hz and $f_3 = 300$ Hz. In Fig. 2b, energy assigned for $s_2[n]$ is at frequencies $f_2 = 200$ Hz and $f_4 = 300$ Hz. Combining the minimized TF-spectra with the corresponding TF-bin phases gives us the separated source. In an anechoic model, the sources in the mixture suffer relative delays. An arbitrary time domain mixture $x[n - \delta]$, incurring a delay of δ samples is represented in frequency domain as $\mathbf{X} : x[n - \delta] \longmapsto \mathbf{X}[k]e^{-j\Omega_k\delta}$. Sources $s_1[n]$ and $s_2[n]$ in mixture $x_2[n]$ undergo a delay of δ_1 and δ_2 in samples, respectively. Then the expressions for the frequency-attenuation matrix $\mathbf{A}$ and its decomposition into spectral $\mathbf{W}$ and attenuation $\mathbf{H}$ components are derived below:

$$\mathbf{A_1} = \left|\mathbf{X}_1[k] - g\mathbf{X}_2[k]e^{-j\Omega_k\delta}\right| = \Big|\delta[k - f_1] + \delta[k - f_3] + \delta[k - f_2] + \delta(k - f_4) - g\big[\alpha_1 e^{-j\Omega_k\delta_1}\big(\delta[k - f_1] + \delta[k - f_3]\big) + \alpha_2 e^{-j\Omega_k\delta_2}\big(\delta[k - f_2] + \delta[k - f_4]\big)\big]\Big| \quad (11)$$

Solving Eqn. 11, we get the below:

$$\mathbf{A_1} = \left|\Big(\delta[k - f_1] + \delta[k - f_3]\Big)\Big(1 - g\alpha_1 e^{-j\Omega_k\delta_1}\Big) + \Big(\delta[k - f_2] + \delta[k - f_4]\Big)\Big(1 - g\alpha_2 e^{-j\Omega_k\delta_2}\Big)\right| \quad (12)$$



(a) Redress $s_1[n]$

(b) Redress $s_2[n]$

(c) D-Redress(Cancel) $s_1[n]$

(d) D-Redress(Cancel) $s_2[n]$

Fig. 2. Recovered TF-spectra $\mathbf{W}$: The blue line depicts the TF-spectrum of the original signal $s_j[n]$. The red line signifies the TF-spectrum of the estimated version of that signal. Two synthetic sources overlap at $k = 300$ Hz. **Redress**: Figure (a) illustrates that the source $s_1[n]$ has energy at $k = 100$Hz and $k = 300$Hz. In figure (b), the energy assigned to source $s_2[n]$ is at $k = 200$Hz and $k = 300$Hz. **D-Redress(Cancel)**: The figures (c) and (d) shows similar performance to the Redress. The difference is that a minor power assignment to $s_1[n]$ at 200 Hz in figure (c) is observed.

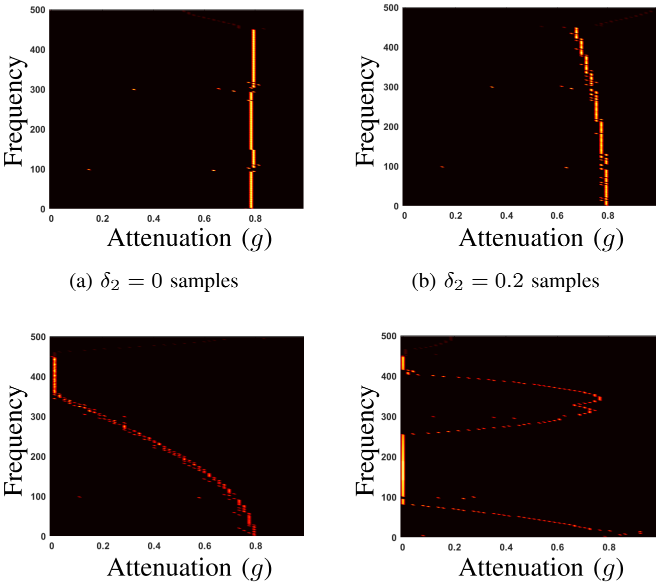
In a similar way, $\mathbf{A_2} = \left|\mathbf{X}_2[k]e^{-j\Omega_k\delta} - g\mathbf{X}_1[k]\right|$ can be represented as:

$$\mathbf{A_2} = \left|\Big(\delta[k - f_1] + \delta[k - f_3]\Big)\Big(\alpha_1 e^{-j\Omega_k\delta_1} - g\Big) + \Big(\delta[k - f_2] + \delta[k - f_4]\Big)\Big(\alpha_2 e^{-j\Omega_k\delta_2} - g\Big)\right| \quad (13)$$

From Eqn. 12 and Eqn. 13 we get the null locations on matrix $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$ on frequencies f_1, f_2, f_3, f_4 at gains $g = \frac{1}{\alpha_1 e^{-j\Omega_k\delta_1}}$ or $g = \alpha_1 e^{-j\Omega_k\delta_1}$ and at $g = \frac{1}{\alpha_2 e^{-j\Omega_k\delta_2}}$ or $g = \alpha_2 e^{-j\Omega_k\delta_2}$, subject to $g \leq 1$, and $j = \sqrt{-1}$. The location of null peaks are characterized by the quantity $e^{-j\Omega_k\delta_j}$. Here the subscript $j = 1, 2$ is the source index. Estimating the TF-spectra for anechoic mixtures is not as straight forward as the instantaneous scenario. Therefore, we devise a novel technique D-Redress which mitigates this issue.

IV. D-Redress

We have extended the Redress by two methods: firstly, using the attenuation component $\mathbf{H}$ for instantaneous mixtures, Eqn. 9, but cancelling-out the delay in the mixture. This is known as D-Redress (Cancel). The other method is adapting $\mathbf{H}$ to a complex attenuation component as depicted by Eqn. 12 and Eqn. 13, called D-Redress (Complex). **D-Redress(Cancel)**: Redress groups the null-peaks of the target source at an appropriate g location based upon its position



(a) $\delta_2 = 0$ samples

(b) $\delta_2 = 0.2$ samples

(c) $\delta_2 = 0.7$ samples

(d) $\delta_2 = 5$ samples

Fig. 3. The position of null-peaks on the frequency-attenuation matrix $\mathbf{A}$ changes with delays δ_j. Figure (a) illustrates sources with no delays, the peaks are clustered at $g \approx 0.8 = \alpha_2$. Figure (b) depicts that a delay of 0.2 samples causes the null-peaks bend. Figure (c) shows that a delay of 0.7 samples shifts the peaks to the extreme left. Figure (d) illustrates that a delay of 5 samples makes it difficult to determine the null-peak locations.

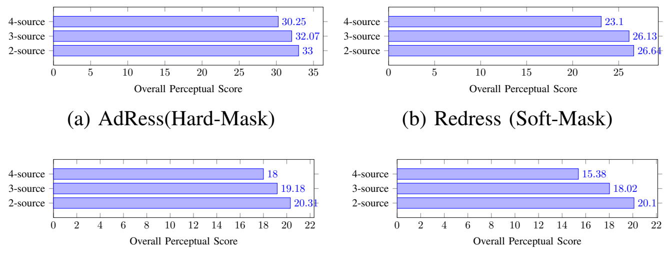
in the mixture, provided the sources frequencies are non-overlapping. Let the ground-truth location be $\alpha_2 = 0.8$ in our mixture. Fig. 3a depicts that with no delay, the null-peaks of $s_2[n]$ are grouped at $g \approx 0.8 = \alpha_2$. A delay of $\delta = 0.2$ samples causes the alignment of this grouping to bend to the left, as shown in Fig. 3b. For $\delta = 0.7$ samples, this deviation gets more pronounced and the peaks start to get shifted to the extreme left, as depicted in Fig. 3c. Lastly, in Fig. 3d, at $\delta = 5$ samples, the null-peaks get entirely dislocated. Thus, determining the location of the null-peaks is difficult. This makes $\mathbf{A}$ hard to use for instantaneous de-mixing. The D-Redress(Cancel) method implements cancellation of relative delays in the mixture [12]. If δ_2 is cancelled-out in the mixture $\mathbf{X_2}$, that is $\mathbf{X_2} = e^{+j\Omega_k\delta_2} \odot \mathbf{X_2}$, then the disorganized null-peak shall again be grouped at $g = 0.8 = \alpha_2$. This is equivalent to the no-delay case as depicted in Fig. 3a. More appropriately the delay in $\mathbf{X_2}$ has been cancelled-out, the probability of obtaining $\mathbf{A}$ similar to the instantaneous mixing scenario is increased. Let us have a 2-source mixture with sources $s_1[n]$ and $s_2[n]$ and delays δ_1 and δ_2. Cancelling δ_2 and then running the same quadratic program update with $\mathbf{H}$ (of Eqn. 9 and Eqn. 10) gives an estimate of the TF-spectrum of $s_2[n]$. Fig. 2c and Fig. 2d depict the TF spectral extraction of $s_1[n]$ and $s_2[n]$, respectively. In the extracted TF spectrum, we assume that the power of one source dominates over the other sources. Empirically for a 2-source case, if we cancel $\mathbf{X_2}$ by any one of the delay, the corresponding source is separated. Let us have a set of delays in samples $\delta_j = \delta_1, \delta_2, \delta_3, \delta_4$. For a 3-source case, the best human-ear intelligibility of the separated sources holds when $\mathbf{X_2}$ is cancelled-out by δ_2. For a 4-source case, the most intelligible speech is extracted when cancelled by $\delta = \frac{\delta_2+\delta_3}{2}$ samples. **D-Redress(Complex)**: In this approach, we adapt the set of discrete frequencies $\Omega = \{-\pi, \ldots, \ldots, +\pi\}$, where $|\Omega_k| \leq \pi \in \Omega$, to build the attenuation component $\mathbf{H}$ for anechoic mixtures. In our experiment, the size of g is set to be $G = 100$. We divide Ω into 100 equal parts, then adapt it in Eqn. 12 and Eqn. 13. This formulates the anechoic $\mathbf{H}$ similar to Eqn. 9 and Eqn. 10 . We repeat the same quadratic program update to extract the magnitude-spectra $\mathbf{W}$.

V. Experiments and Results

Our proposed D-Redress algorithm is compared with three other techniques, namely AdRess, D-AdRess and Redress. The quality of the separated speech utterances are evaluated using BSS Eval [17] and PEASS [18] toolkits. We have randomly selected the utterances from a total number of 25200 files of the TIMIT corpus [19]. Each speech utterance is sampled at 16 kHz. In our experiments, we have considered a 1024 sample Hamming window for a 50% overlap. Our evaluations are based upon 1000 Monte Carlo trials. We observe that: (1) Mean reconstruction Source-to-Distortion Ratio (SDR) of Redress exceeds D-Redress(Cancel) and D-Redress(Complex) by 0.2 dB and 0.5 dB respectively. (2) Source-to-Artifact Ratio (SAR) of Redress exceeds AdRess by 0.5 dB. (3) Mean reconstruction Source-to-Interference Ratio (SIR) of Redress, D-Redress(Cancel) and D-Redress(Complex) is less than AdRess by 2 dB, 3.5 dB and 5 dB, respectively.

The quality of separated utterances decrease as SS move from instantaneous to anechoic mixing models. Higher values of SDR, SAR and SIR signify better performance. We have considered up to four sources ($J = 4$) in the mixture. The attenuation coefficients α_j's are chosen in a such a way that they are distant from one another. Otherwise, the null-peaks formed on $\mathbf{A}$, Fig. 3a, shall be very close in the g axis. This may make extraction of spectral-power of the desired source from the appropriate sub-portion of $\mathbf{A}$ difficult. Energy from the neighbouring sources may get included or "leaked" into the desired utterance. In our experiments, for a 4-source scenario, the attenuation values are chosen as: $\alpha_1 = 0.20, \alpha_2 = 0.45, \alpha_3 = 0.70, \alpha_4 = 0.98$. A 3-source scenario consists of $\alpha_1 = 0.20, \alpha_2 = 0.55, \alpha_3 = 0.85$. The 2-source case consists of $\alpha_1 = 0.20, \alpha_2 = 0.80$. The delays considered (in samples) are $\delta_1 = 10, \delta_2 = 7, \delta_3 = 4, \delta_4 = 1$ for a 4-source mixture. We have also considered other values of delays too. The algorithms are coded in Matlab R2021a. Since the Redress is a soft-masking technique, the reconstructed TF spectra shall have negligible number of holes. Unlike AdRess, the Redress assigns the TF bins some spectral power implying better SAR as depicted in Fig. **??**. Fewer artifacts implies musical-noise. The drawback here is that the interference in the separated utterances gets more.

Fig. **??** depicts that the SIR for Redress is less than AdRess by 1 dB. The PEASS toolkit [18] is used to evaluate the



(a) AdRess(Hard-Mask)

(b) Redress (Soft-Mask)

(c) D-Redress (Cancel)

(d) D-Redress (Complex)

Fig. 4. PEASS Score: Overall Perceptual Score (OPS) is measured. **AdRess**: Figure (a) illustrates that the mean reconstruction OPS is 31.6. **Redress**: Figure (b) depicts the mean reconstruction OPS is 25.28. **D-Redress(Cancel)**: Figure (c) shows the mean reconstruction OPS is 19.18. **D-Redress(Complex)**: Figure (d) illustrates that the mean reconstruction OPS is 17.84.

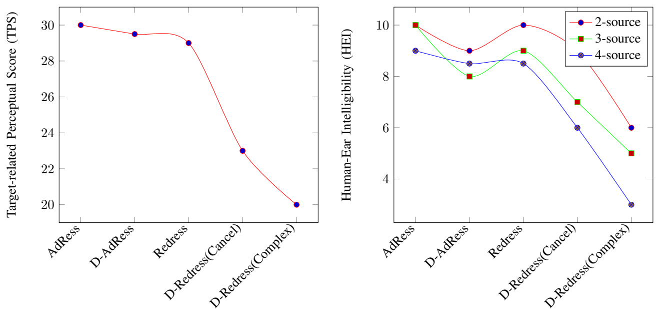
perceptual-score. This score is reduced for anechoic mixing scenario. The PEASS-score [18] of the estimated source, $\hat{\mathbf{s}_j}$, is measured in a scale within $1 - 100$. Higher values signify better performance. Let us have J number of speech signals in the mixture index by j. PEASS seeks to split the distortion between the estimated $\hat{\mathbf{s}_j}$ and the target-original $\mathbf{s}_j$ into given as: $\hat{\mathbf{s}_j} - \mathbf{s}_j = \mathbf{e}_j^{\text{target}} + \mathbf{e}_j^{\text{interference}} + \mathbf{e}_j^{\text{noise}}$. Fig. 4 shows, the Overall Perceptual Score (OPS) of the Redress decreases as extended to anechoic scenario. The OPS describes the perceptual-similarity measure (PSM) [20], between the original and the estimated as perceived by a human ear. The mean reconstruction OPS of the Redress surpasses its anechoic variants, by 6.1 for D-Redress (Cancel) and 7.84 for D-Redress (Complex). The Target related-Perceptual-Score (TPS) means PSM$\left(\hat{\mathbf{s}_j}, \hat{\mathbf{s}_j} - \mathbf{e}_j^{\text{target}}\right)$. This is the perceptual similarity measure between the estimated utterance $\hat{s_j}$ with itself, minus the target-distortion component $\mathbf{e}_j^{\text{target}}$. Fig. 5a shows that the TPS of Redress is 29. It drops to 22 for D-Redress(Cancel) and to 20 for D-Redress(Complex). Lastly, we perform human-ear intelligibility, a study involving 5 participants, they were asked to access each sound and mark them somewhere in between $1 - 10$. The perceived utterance's intelligibility was measured, bench-marked against a set of reference values as follows:

- Completely similar-10
- Sound Quality deteriorates but very intelligible-7
- Sound Quality bad, but intelligible-4
- Sounds unintelligible-1

Fig. 5b depicts Human-Ear Intelligibility (HEI) of the separated speech utterance decreases with increase in number of sources in the mixture.

VI. Conclusion and Future work

The soft-mask Redress reduces the artifacts in the separated speech. This gives the Redress a higher SAR compared to the AdRess technique. On the other hand, the Redress introduces more interference to the separated speech utterances compared to AdRess. Consequently, the SIR of Redress is than AdRess. Perceptual-quality, distortions, artifacts, interference and human-ear intelligibility of the utterances separated by the Redress were reduced when extended to anechoic scenarios. An increase in the number of constituent sources in the mixture also decreases these metrics. Perceptually, D-Redress(Cancel) is more satisfactory than D-Redress(Complex). In our evaluations, we have considered small delays up to 10 samples. In future, we shall evaluate how close the neighbouring sources can be on the frequency-attenuation matrix, while still achieving good separation. We shall also consider how to adapt Redress to big delay scenarios that might occur in real-world sensor networks.



(a) Target-Perceptual Score

(b) Human-Ear Intelligibility

Fig. 5. Figure(a) depicts that the Target-Perceptual Score (TPS) of instantaneous mixtures have more overall scores their anechoic counterparts. Soft-mask D-Redress(Cancel) and D-Redress(Complex) is less than Redress by 6 and 10 respectively. Hard-mask D-AdRess for anechoic mixing scenarios is less than AdRess by 1. Figure (b) illustrates that the Human-Ear Intelligibility (HEI) of the separated speeches become less intelligible with increase in number of sources in the mixture.

Acknowledgment

This paper has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant number 18/CRT/6222. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/SIRG/3459.

References

[1] R. H. Roy III and T. Kailath, "ESPRIT-Estimation of Signal Parameters via Rotational Invariance Techniques," Optical Engineering, vol. 29, no. 4, pp. 296–313, 1990.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Sig. Proc., vol. 52, no. 7, pp. 1830–1847, July 2004.

[3] F. Abrard and Y. Deville, "A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," Sig. Proc., vol. 85, no. 7, pp. 1389–1403, 2005.

[4] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in International Conf. ICA and Sig. Sep. Springer, 2006, pp. 536–543.

[5] T. Melia, S. Rickard, and C. Fearon, "Histogram-based blind source separation of more sources than sensors using a DUET-ESPRIT technique," in 13th EUSIPCO, 2005, pp. 1–4.

[6] R. de Fréin and S. T. Rickard, "Power-weighted divergences for relative attenuation and delay estimation," IEEE Sig. Proc. Let., Nov 2016.

[7] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," IEEE Tran. on Neural Networks, 1999.

[8] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," Neural Computation, 2001.

[9] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. on Sig. Proc., 2004.

[10] R. de Fréin, “Reformulating the binary masking approach of Adress as soft masking,” Electronics, vol. 9, no. 9, p. 1373, 2020.

[11] D. Barry, B. Lawlor, and E. Coyle, “Sound source separation: Azimuth discrimination and resynthesis,” in 7th DAFX 04, 2004.

[12] S. Bagchi and R. de Fréin, “Extending instantaneous de-mixing algorithms to anechoic mixtures,” in IEEE ISSC, 2021, pp. 1–6.

[13] R. de Fréin and S. T. Rickard, “The synchronized short-time-Fourier-transform: Properties and definitions for multichannel source separation,” IEEE Trans. Sig. Proc., vol. 59, no. 1, pp. 91–103, Jan 2011.

[14] K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” Speech Commun., 2011.

[15] R. de Fréin, “Remedying sound source separation via azimuth discrimination and re-synthesis,” in IEEE ISSC, 2020, pp. 1–6.

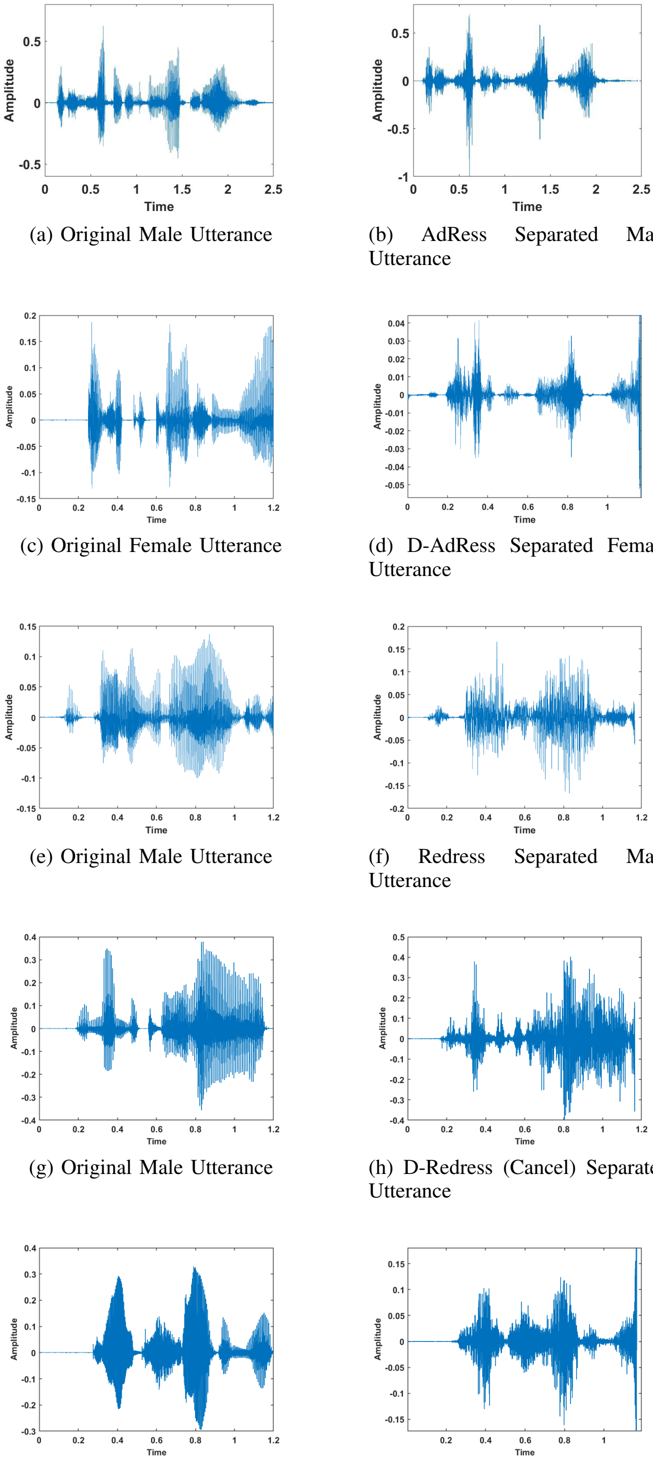
[16] D. Seung and L. Lee, “Algorithms for non-negative matrix factorization,” Adv. Neural Inf. Process. Syst., vol. 13, pp. 556–562, 2001.

[17] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” IEEE Trans. on Aud., Speech and Lang. Proc., vol. 14, no. 4, pp. 1462–1469, 2006.

[18] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, “Subjective and objective quality assessment of audio source separation,” IEEE Trans. Audio Speech Lang. Process., vol. 19, no. 7, pp. 2046–2057, 2011.

[19] J. S. Garofolo, “TIMIT acoustic phonetic continuous speech corpus,” LDC, 1993.

[20] R. Huber and B. Kollmeier, “PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception,” IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 6, pp. 1902–1911, 2006.



(a) Original Male Utterance

(b) AdRess Separated Male Utterance

(c) Original Female Utterance

(d) D-AdRess Separated Female Utterance

(e) Original Male Utterance

(f) Redress Separated Male Utterance

(g) Original Male Utterance

(h) D-Redress (Cancel) Separated Utterance

(i) Original Female Utterance

(j) D-Redress(Complex) Separated Utterance

Fig. 6. Column1: Original utterances (TIMIT database). Column2: Estimated utterances from a mixture of 2-sources. Top-Down order: AdRess, D-AdRess, Redress, D-Redress(Cancel), D-Redress(Complex); As we extend the instantaneous algorithms to anechoic scenarios, incongruities in the temporal structures of the separated utterances are observed. This indicates increased distortion and interference. Moving from instantaneous to anechoic scenarios decreases the intelligibility of the separated utterances. Increasing the number of sources to four in the mixture decreases the quality and intelligibility.