Implementation and Assessment of Joint Source Separation and Dereverberation

MOFFAT, DJ; reiss, JD; Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)

Implementation and Assessment of Joint Source Separation and Dereverberation

David Moffat, and Joshua D. Reiss
Center for Digital Music, Queen Mary University of London
me@davemoffat.com

July 6, 2016

Abstract

Reverberation is known to introduce difficulties in audio source separation, and reverse engineering independent sources from a convolutive mixture is one of the toughest challenges within blind source separation. This paper proposes two novel methods that combine dereverberation work with microphone interference reduction. The results are evaluated objectively using the BSS Eval toolbox and Reverb Workshop Evaluation Toolbox, relative to the effectiveness of the dereverberation and source separation. Both proposed methods show improvements on the existing dereverberation technique used. However, this has a negative impact on the source separation, as has also been seen in other work. An explanation for this negative impact and alternative approaches to avoid this situation are proposed.

1 Introduction

Microphones are often used in music and speech reinforcement and audio recording, however there are some fundamental issues whenever more than a single microphone is used to record a scene. Whenever more than a single source is being recorded simultaneously, there will be some interference between the sources, where more than a single source will be picked up in each source microphone. This interference will reduce the ability to distinguish each individual source and cause comb filtering [1].

In the basic two microphone, two source situation, each microphone has a target source with a direct path, and then an interfering source path, as presented in Figure 1. In the ideal case, the direct path is the only source we want to be captured by the microphone. Where we have a microphone signal x_m and a source signal s_m, x_m can be defined as

$$x_m[n] = h_{s,m}[n] * s_m[n] \tag{1}$$

where $h_{s,m}$ is the Acoustic Impulse Response (AIR) from the source s to the microphone m, and $*$ denotes the convolution operation. For the purposes of

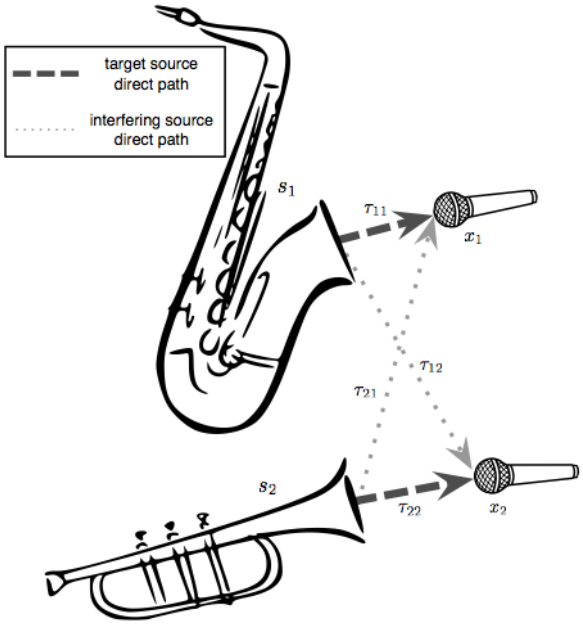

Figure 1: Real world first order of interference case, taken from [2]

this paper, the following notation will be used. A system input signal is defined as $\mathbf{x}_m[n]$, for the mth channel such that $0 \leq m \leq M$, where M is the total number of microphones. $\mathbf{X}_m[k]$ is defined as $\mathcal{F}(\mathbf{x}_m[n])$, where $\mathcal{F}$ denotes the Fourier Transform. As this paper only considered the fully determined case, the total number of microphones is equal to the number of sources. Further to this, we will only consider the close microphone situation within this paper, so the following assumptions are made. Each mic is intended to pick up one source. A mic will pick up the intended source much stronger than any other. The intended source will have very little reverb, since it is mainly received in the direct path.

Existing work attempts to reduce the interference between the two signals [2, 3, 4]. However, reverberation causes substantial issues in current research into microphone interference reduction or blind source separation [2, 5, 6, 7, 8].

This paper extends current microphone interference reduction work by [2] with the application of a blind dereverberation method by [9]. This work is applicable both to recorded and live sound environments, any environment where more than a single microphone is being used to record multiple sources.

This paper will discuss the use of microphone interference cancellation using Crosstalk-Resistant Adaptive Noise Canceller (CTRANC) and the basics of reverberation and its removal. Section 2 presents the novel work undertaken as part of this project, in which approaches to joint dereverberation and source

separation, based on adaptation and extension of the CTRANC algorithm, are described. Section 3 presents the results and the evaluation of these. Section 4 provides a conclusion in which the principal challenges and further work will be discussed.

1.1 Microphone Interference Reduction

There have been many different approaches to blind source separation and noise reduction, however some of the most recent work, focusing on the real time live sound environment has produced exciting new work. Within audio engineering, the most recent work on interference reduction by [10], has produced some effective results by the use of adaptive Weiner filters called Selective Frequency Domain CTRANC (selFDCTRANC), and it is the intention to extend this work. Interfering microphone x_l is defined as

$$\mathbf{X}'_l[k] = diag\left(\mathcal{F}[x_l[kN], \ldots, x_l[kN+N-1], 0, \ldots, 0]^T\right) \tag{2}$$

where $x_l[n]$ denotes in input audio signal for sample n, where l is the channel of interference and N is the length of the adaptive Weiner filter. The previous and current interference channels are summed together to produce the interfering channel vector

$$\mathbf{X}_l[k] = \mathbf{X}'_l[k] + \mathbf{J}\mathbf{X}'_l[k-1] \tag{3}$$

where, due to the overlap and add conditions $\mathbf{J} = diag[1, -1, 1, -1, 1, -1 \ldots, 1]$. Filter weights are calculated as

$$\phi[k] = \sum_{l=1, l\neq m}^{L} \mathbf{X}_l[k] + \mathbf{W}_{lm}[k] \tag{4}$$

and the output is updated by the following

$$\hat{\mathbf{x}}_m[k] = \mathbf{x}_m[k] - \mathcal{F}^{-1}\phi[k] \tag{5}$$

Interfering microphones are then updated in the frequency domain by the filter update equation:

$$\mathbf{W}_{lm}[k+1] = \mathbf{W}_{lm}[k] + \mathcal{F}^{-1}\mu[k]\mathbf{X_l}[k]^H\hat{\mathbf{X}}_m[k] \tag{6}$$

where

$$\mu[k] = \mu \cdot diag(T^{-1}[k]) \tag{7}$$

where μ denotes the frequency dependent step size. The forgetting factor γ is then applied as such:

$$T[k] = \gamma T[k-1] + (1-\gamma)|X_l[k]|^2 \tag{8}$$

1.2 Dereverberation

Dereverberation is the process of removing or reducing reverberation from a signal. The most common case investigated is blind dereverberation. This assumes only knowledge of the recorded signal, $x_m[n]$, to calculate $s_m[n]$, either

by estimating $h_{s,m}[n]$ or by estimating $s_m[n]$ directly. Dereverberation can be broken down into three different methods Beamforming, Spectral Enhancement and Blind Deconvolution.

We chose an adaptive dereverberation method based on statistical modelling that is a form of spectral enhancement [9]. It follows a similar adaptive filter design as [2]. The two methods share a number of similarities, as they are both real time capable, implementing adaptive filters to clean up a signal, include some short term signal memory and convergence variable. The purpose of this method is to design a gain function $\mathbf{G}[k]$ such that

$$\hat{\mathbf{X}}[k] = \mathbf{G}[k]\mathbf{X}[k] \tag{9}$$

To perform this gain calculation, the power spectrum is taken over short term and long term moving averages which are denoted by $\mathbf{R}_1[l,k]$ and $\mathbf{R}_2[k]$ respectfully.

$$\mathbf{R}_1[k] = (1-\alpha_1)\mathbf{P}[k] + \alpha_1\mathbf{R}_1[k-1] \tag{10}$$

$$\mathbf{R}_2[k] = (1-\alpha_2)\mathbf{P}[k] + \alpha_2\mathbf{R}_2[k-1] \tag{11}$$

Where $\mathbf{P}[k] = |\mathbf{X}[k]|^2$ and $0 < \alpha_1 < \alpha_2 < 1$. From this we can calculate a gain function as:

$$\mathbf{G}[k] = \begin{cases} 1, & \dfrac{\mathbf{R}_1[k]}{\mathbf{R}_2[k]} \geq 1 \\ \dfrac{\mathbf{R}_1[k]}{\mathbf{R}_2[k]} & \text{otherwise} \end{cases} \tag{12}$$

It is assumed that the effect of reverberation can be represented by the Modulation Transfer Function (MTF) [9] As such, the MTF represents the difference between the input sound and the output sound [9]. The MTF $A[f_m]$, where f_m is the modulation frequency, can be estimated as:

$$A[f_m] = \frac{1}{\sqrt{1 + \left(2\pi f_m \dfrac{RT_{60}}{6\log_e(10)}\right)^2}} \tag{13}$$

We can also compute the frequency response $H[f_m]$ of our calculated dereverberation gain $G[k]$ as:

$$H[f_m] = \frac{1-\alpha_1}{1-\alpha_2}\frac{1-\alpha_2 e^{-j2\pi f_m}}{1-\alpha_1 e^{-j2\pi f_m}} \tag{14}$$

To estimate the forgetting factors α_1 and α_2, we assume perfect dereverberation can occur, such that

$$|H[f_m]|A[f_m] = 1 \tag{15}$$

So we can estimate the forgetting factors α_1 and α_2 to reduce the sum of squares error function

$$E = \frac{1}{2}\sum_{m=1}^{M}\left(1 - |H[f_m]|A[f_m]\right)^2 \tag{16}$$

This allows an estimation of the smoothing constants using the steepest descent method, to deduce the error with iterative equations

$$\alpha_1[i+1] = \alpha_1[i] - \lambda_1 \frac{\partial E}{\partial \alpha_1} \tag{17}$$

$$\alpha_2[i+1] = \alpha_2[i] - \lambda_2 \frac{\partial E}{\partial \alpha_2} \tag{18}$$

$$\frac{\partial E}{\partial \alpha_1} = -\frac{1+\alpha_1}{1-\alpha_1} \sum_{m=1}^{M} \frac{1-\cos(2\pi f_m) B[f_m]}{|1-\alpha_1 e^{-j2\pi f_m}|^2} \tag{19}$$

$$\frac{\partial E}{\partial \alpha_2} = \frac{1+\alpha_2}{1-\alpha_2} \sum_{m=1}^{M} \frac{1-\cos(2\pi f_m) B[f_m]}{|1-\alpha_2 e^{-j2\pi f_m}|^2} \tag{20}$$

$$B[f_m] = |H[f_m]|A[f_m]\left(|H[f_m]|A[f_m] - 1\right) \tag{21}$$

2 Joint Microphone Interference Reduction and Dereverberation

A two stage approach is implemented in which dereverberation is applied to an audio signal prior to the application of any interference reduction. As can be seen from Figure 2, dereverberation is applied to each microphone channel independently and each signal is passed as the input to the interference reduction algorithm. The implementation resembles the dereverberation method from [9], where the second half of the algorithm represents the microphone interference reduction method from [2].

A combined dereverberation and microphone interference reduction method is also proposed. Existing work is combined and this will be referred to as selFDCTRANC with Dereverberation (selFDCTRANCD). The dereverberation and interference reduction occur in parallel, as shown in Figure 3. The input signal is passed to both the dereverberation and the interference reduction aspects of the system, with the interference reduction filter being applied to the dereverberant signal. In order to combine the dereverberation with the microphone interference cancellation algorithm, Equation (5) has now become

$$\hat{\mathbf{x}}_m[k] = \tilde{\mathbf{x}}_m[k] - \mathcal{F}^{-1}\phi[k] \tag{22}$$

where $\tilde{\mathbf{x}}_m[k]$ represents the dereverberant input signal in time domain, such that

$$\tilde{\mathbf{x}}_m[k] = \mathcal{F}^{-1}\left(\tilde{\mathbf{X}}_m[k]\right) \tag{23}$$

$$\tilde{\mathbf{X}}_m[k] = \mathbf{G}_m[k]\mathbf{X}_m[k] \tag{24}$$

3 Results

To simulate a mixing environment, a multitrack from the Open Multitrack Testbed [11], consisting of eight sources, was used. The eight sources were

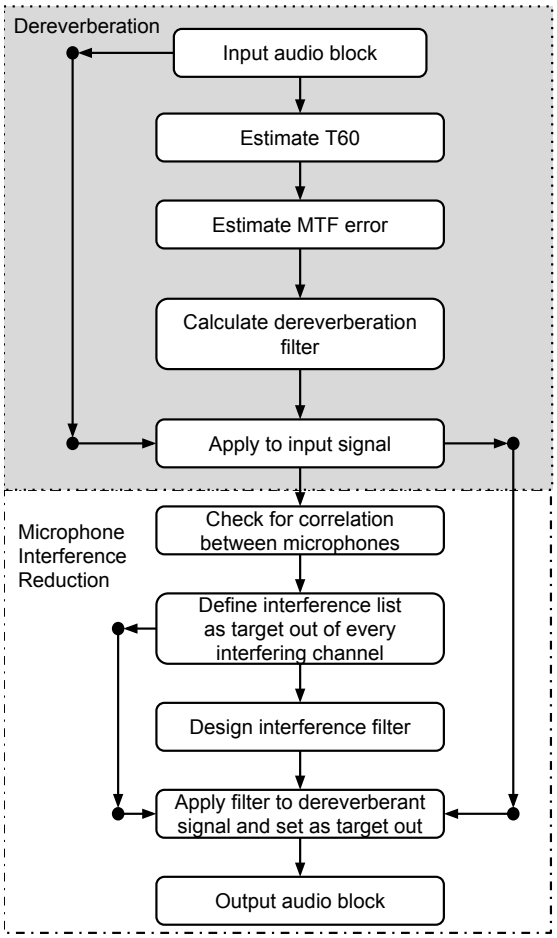

Figure 2: Flow Diagram of the Two Stage Method

spaced around the room and simulation impulse responses were generated with the Room Impulse Response Generator [12]. The room is presented in Figure 4. These were then combined to produce a simulation of eight convolutive mixtures, each representing a single microphone. The simulation was then processed with the following four methods.

- selFDCTRANCD
- Two Stage Method
- Microphone Interference Reduction [2]
- [9] Method

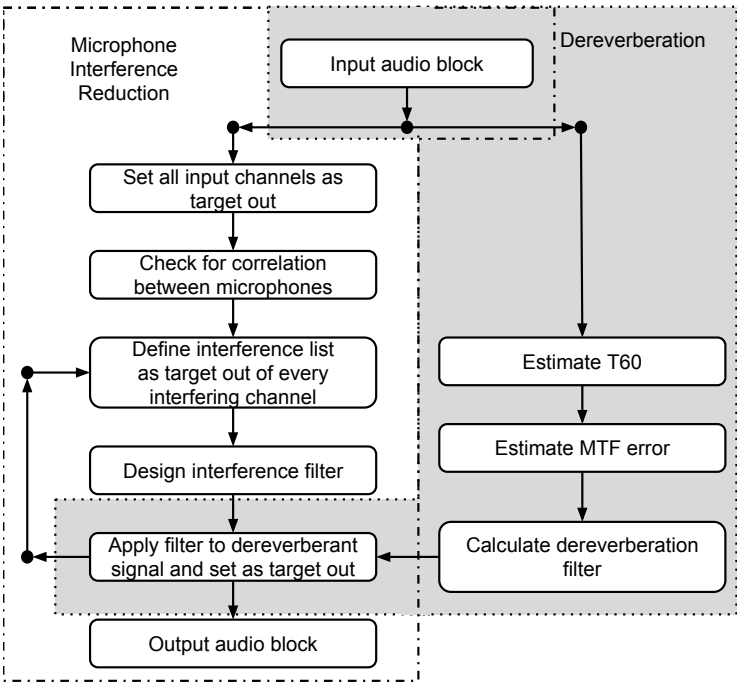

Figure 3: Flow Diagram of selFDCTRANCD

3.1 Evaluation Metrics

The Reverberation Workshop [13] proposed a series of metrics to evaluate dereverberation algorithms. These dereverberation evaluation methods are Cepstral Distance (CD), Log Likelihood Ratio (LLR), Speech-to-Reverberation Modulation Energy Ratio (SRMR) and the frequency weighted Segmental Signal to Noise Ratio (fwSNRsig), as recommended by [14]. For the CD and LLR, a lower value represents a higher quality signal, where as with SRMR and fwSNRsig, a higher number represents a better signal quality.

Within the source separation community, it is generally agreed that the Signal Interference Ratio (SIR) is an effective measure of interference and clarity of a source, and most papers use this in combination with Signal Distortion Ratio (SDR) and Signal Artifact Ratio (SAR). These methods are all presented in the BSS Eval toolbox [15]. Evaluation of source separation will be performed with the BSS Eval toolbox [15], provided in MATLAB, as this toolbox is one of the most used toolboxes within the source separation community. The work from [2] was evaluated using the BSS toolbox, and so for reasonable comparison, evaluation will follow the same metric. For SAR, SIR and SDR, a higher number represents a higher quality signal.

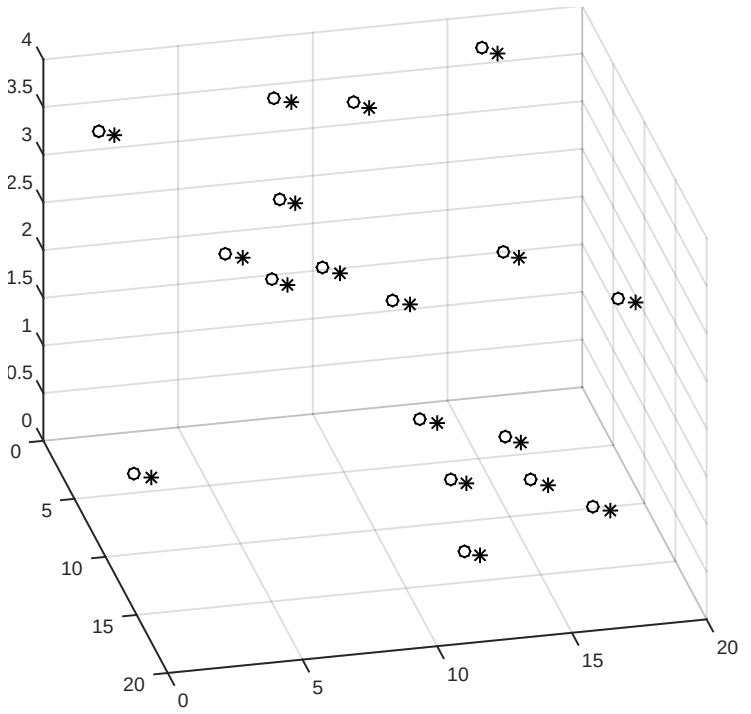

Figure 4: Visualisation of Room Simulation. Room Size 20m x 20m x 4m. (o = Source, $*$ = Microphone)

3.2 Dereverberation Results

The simulation environment was evaluated with the Dereverberation Evaluation Framework [14]. The framework was used to evaluate the reverberant quality of the original simulation microphone, the [9] method, the two stage method, as proposed in Section 2, and the selFDCTRANCD as proposed in Section 2. The results of this evaluation can be seen in Figures 5 to 8. Definitions and further explanation of the evaluation metrics are presented in [14]. It can be see from Figure 5 that the CD of the original input microphone signal is improved by the methods proposed in this paper. The [9] method performs poorly in these examples. It is expected that the poor results are caused by noise within the signal. Comparing these results with the current state of the art work, the existing real-time implementations of dereverberation algorithms produce a CD of between 3dB and 5dB for reverberation times above 0.5s [13]. Both of these implementations, though not outperforming existing work, perform at a similar standard to many existing dereverberation algorithms.

Figure 6 presents the LLR, and it can be seen that the methods proposed in this paper are better than the original microphone source and the existing [9] method. The state of the art real-time results for LLR fall between 0.25 and

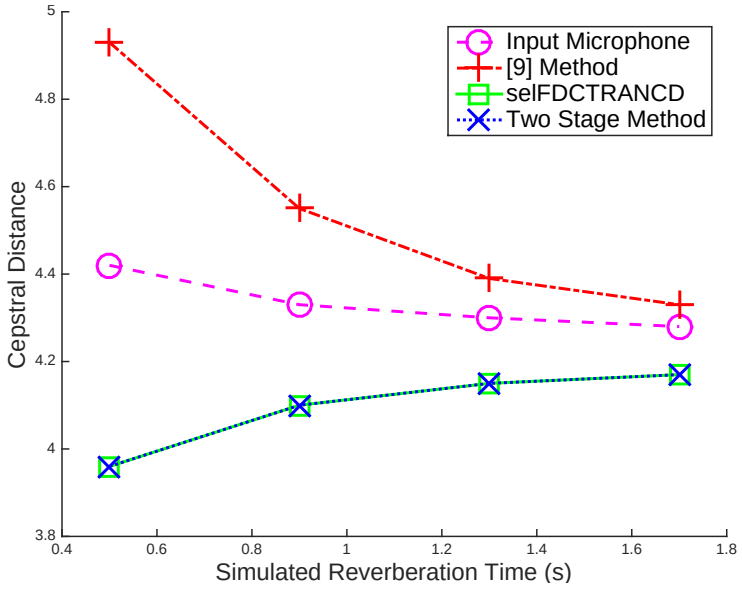

Figure 5: Cepstral Distance Results of Dereverberation Evaluation, selFDC-TRANCD and Two Stage Method produce identical results.

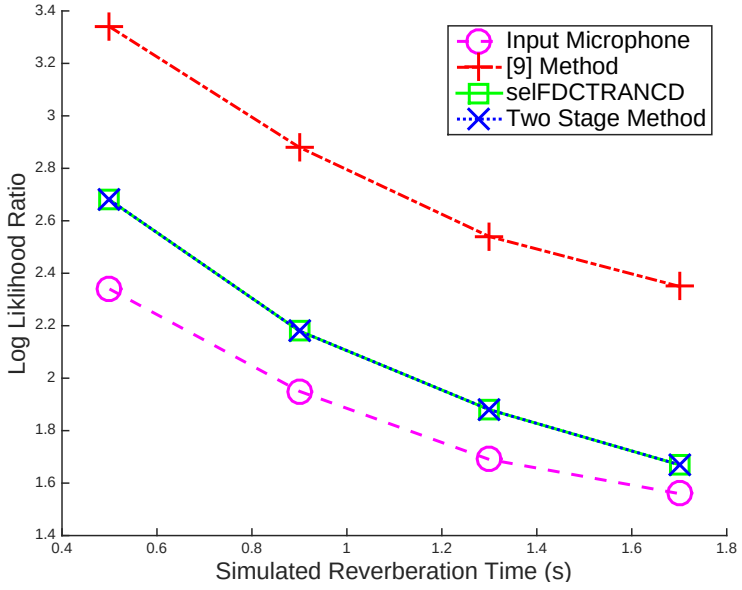

Figure 6: Log Likelihood Ratio Results of Dereverberation Evaluation, selFD-CTRANCD and Two Stage Method produce identical results.

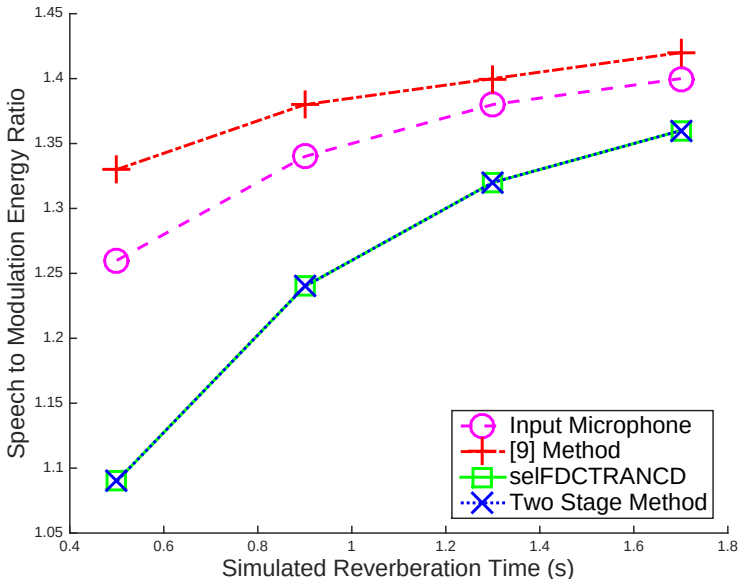

Figure 7: Speech to Reverberation Modulation Energy Ratio Results of Dereverberation Evaluation, selFDCTRANCD and Two Stage Method produce identical results.

0.95, for reverberation times above 0.5s, so though the proposed methods do not perform as effectively as existing work, improvements to the [9] method are presented.

The SRMR, presented in Figure 7 shows that both proposed methods in this paper demonstrate an improvement on the source microphone signal, however [9] clearly outperforms both the selFDCTRANCD method and the two stage method. Existing state of the art, real-time work produces results of SRMR between 3.2 and 8.3, though in these cases the input signal SRMR was between 2.7 and 3.6. It is possible that the lower SRMR is due to the fact that SRMR is designed for speech and that all signals being processed are musical signals.

In Figure 8, it can be seen that [9] makes little improvement to the fwSNRsig. However both the selFDCTRANCD method and two stage method clearly improve the results. It is not reasonable to compare these results with any existing state of the art systems, since the primary measure is based on SNR and existing work looks simply at dereverberation, where significant interference, which will be considered as correlated noise, is added to this system with the reverberation.

3.3 Microphone Interference Reduction Results

The simulation environment, as discussed in Section 3, was also evaluated using the BSS Evaluation MATLAB Toolbox [16, 15]. This toolbox is used to evalu-

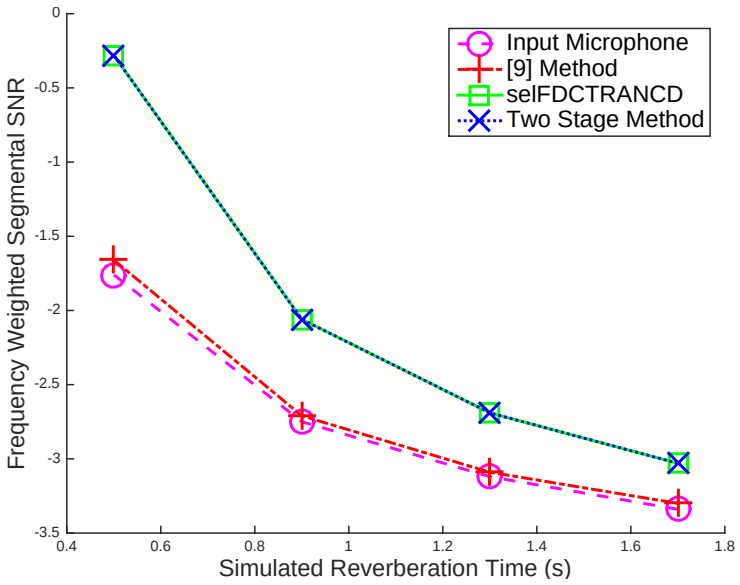

Figure 8: Frequency-Weighted Segmental Signal to Noise Ratio Results of Dereverberation Evaluation, selFDCTRANCD and Two Stage Method produce identical results.

ate the relative source separation that is performed and whether any artifacts or distortion have been introduced into the system. For evaluation purposes, the original source microphone and the selFDCTRANC methods were evaluated alongside both methods proposed in this paper: the selFDCTRANCD and the two stage method. The results are presented in Figures 9 to 11. As these measures are all ratios of the original clean signal with either interference, distortion and artifacts, and as such, results are considered better as the ratios tends to infinity.

Figure 10 presents the SIR results from the BSS evaluation toolbox. It can be seen that all implemented methods are an improvement on the input source microphone. The original selFDCTRANC method outperforms both the selFDCTRANCD and the two stage method, though there is minimal difference, often less than 1.5dB. The selFDCTRANCD method does slightly outperform the two stage method in most cases. The SDR, presented in Figure 9, shows that the selFDCTRANCD and two stage method both show improvements in the input audio signal, however the original selFDCTRANC method performs better than any improvements proposed by this paper. Figure 11 shows that the selFDCTRANC method outperforms both the selFDCTRANCD and two stage method, but that performing any processing will perform better than the original source microphone. It can be seen in both Figure 9 and Figure 10, that a combined dereverberation method can slightly outperform the two stage method, where dereverberation was essentially run as a preprocessing step, par-

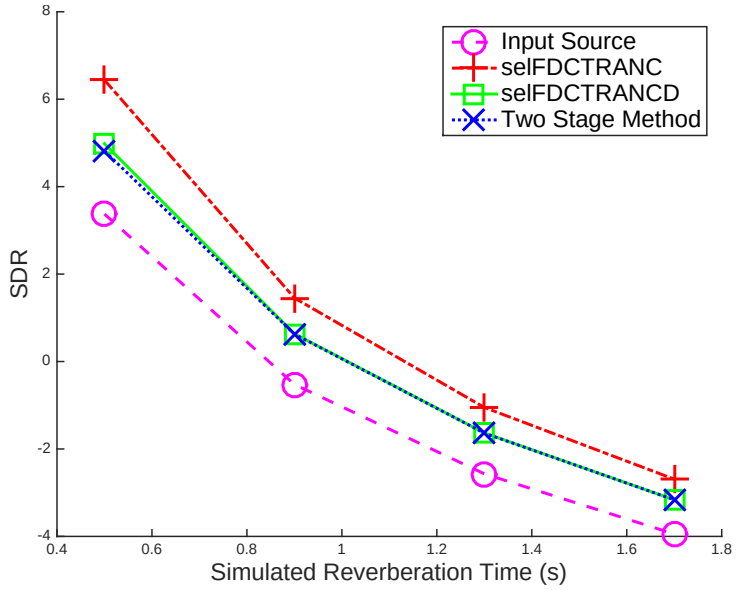

Figure 9: Signal to Distortion Ratio Results of Microphone Interference Reduction Evaluation.

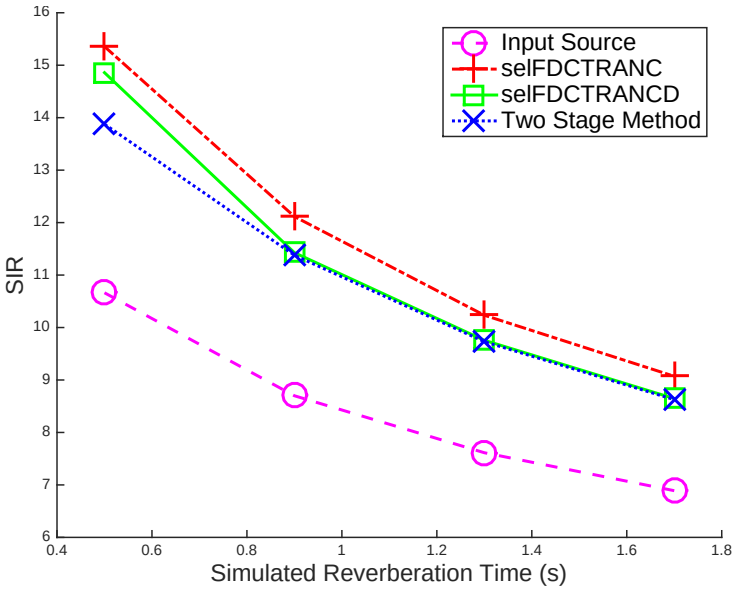

Figure 10: Signal to Interference Ratio Results of Microphone Interference Reduction Evaluation.

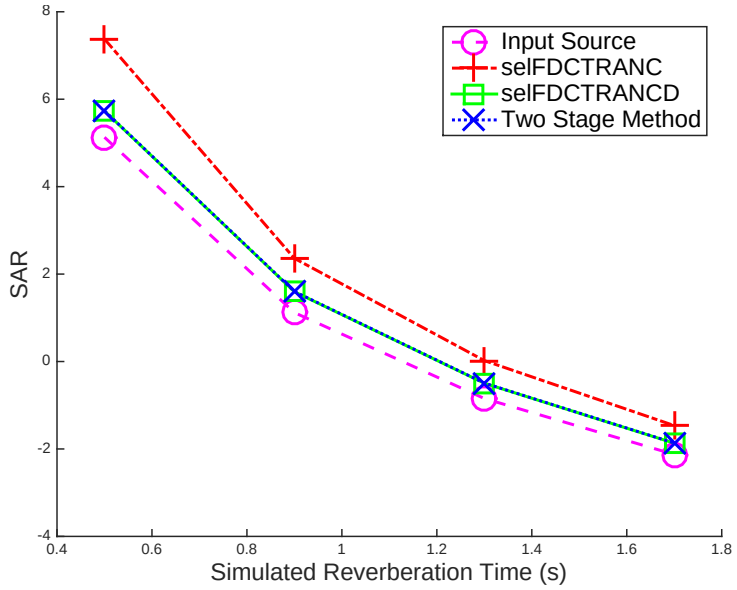

Figure 11: Signal to Artifact Ratio Results of Microphone Interference Reduction Evaluation, selFDCTRANCD and Two Stage Method produce identical results.

ticularly within low reverberation environments.

4 Conclusion

Two novel methods for combining dereverberation with existing work on interference reduction have been proposed, implemented and the results presented. The results demonstrate that the single channel dereverberation has been improved by this multichannel system. However this has had a negative impact on the source separation. It has also been demonstrated that application of source separation can help with dereverberation.

4.1 Discussion

Existing work states that source separation frameworks underperform when working in reverberant environments [2]. Despite this, the results from this paper, along with existing work [17], suggest that implementing spectral enhancement as a form of dereverberation does not improve interference reduction or source separation. However, there is other work where alternative methods of dereverberation, such as beamforming, may improve source separation [18].

Signal intelligibility and clarity issues are only caused by the late reflections of reverberation [19, 20]. As such, current research is focused on removal of

late reverberations within an audio signal. These late reflections will act as correlated noise on a signal, whereas the early reflections will not be perceived as separate from the source signal, so there is little requirement for these to be removed as part of dereverberation. In a multiple source multiple microphone situation, there are first order reflections that are louder as interference in a microphone than in the source microphone. It is proposed that these first few orders of reflections are the cause of difficulties within audio source separation and interference reduction in any convolutive audio mixture.

The assumption based around existing source separation algorithms, is that a direct source arriving at a microphone will always arrive earliest and be the loudest source in a microphone, however it is likely that first order reflections will arrive at an interfering microphone before the source microphone and be louder in the interfering microphone. As such, they first order reflections of an interfering source will remain in a source signal.

4.2 Further Work

This paper has proposed a justification for the lack of source separation improvement through application of dereverberation. This has introduced a new research question as to the cause of source separation issues in convolutive mixtures. Further work is required to uncover the extent to which early reflections cause issues in source separation. The slight improvement in results between the selFDCTRANCD and Two Stage method suggest that there may be justification for producing a combined dereverberation and source separation method, however further work is required in this area of research.

Further work in dereverberation with source separation should focus on the removal or cancellation of the early reflections. Particularly, no form of spectral subtraction is likely capable of removing early reflections from an audio signal, and so will never improve source separation. Beamforming and Blind System Identification can apply early reflection removal, so would be effective research directions, as presented in [21]. Perceptual evaluation of the audio results could also provide some interesting insight into the effectiveness of the proposed methods. Audio is heavily influenced by human perception, and there may be instances where differences between tracks are identified by objective metrics, but imperceivable by listening experts. As such, perceptual evaluation could produce an interesting comparison and more effective evaluation than any available objective measures.

References

[1] G. Ballou, *Handbook for sound engineers*, Focal Press, 2013.

[2] A. Clifford, *Reducing Microphone Artefacts in Live Sound*, Ph.D. thesis, Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University London, 2013.

[3] E. Kokkinis, J. Reiss, and J. Mourjopoulos, "A wiener filter approach to microphone leakage reduction in close-microphone applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2012.

[4] M. Terrell, J. D. Reiss, and M. Sandler, "Automatic noise gate settings for drum recordings containing bleed from secondary sources," *EURASIP Journal on Advances in Signal Processing*, 2010.

[5] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 2.

[6] S. Araki, R. Mukai, et al., "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 2, 2003.

[7] R. Mukai, S. Araki, et al., "Evaluation of separation and dereverberation performance in frequency domain blind source separation," *Acoustical Science and Technology*, vol. 25, no. 2, 2004.

[8] R. Mukai, S. Araki, and S. Makino, "Separation and dereverberation performance of frequency domain blind source separation," in *Proceeding of ICA 2001 Conference*, 2001.

[9] K. Ohtani, T. Komatsu, et al., "Adaptive dereverberation method based on complementary wiener filter and modulation transfer function," *The REVERB workshop*, 2014.

[10] A. Clifford and J. D. Reiss, "Microphone interference reduction in live sound," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, 2011.

[11] B. De Man, M. Mora-Mcginity, et al., "The open multitrack testbed," in *137th Convention of the Audio Engineering Society*, October 2014.

[12] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, 2006.

[13] K. Kinoshita, Ed., *The REVERB workshop*, Firenze Fiera, Florence, May 10, 2014, Workshop in conjunction with 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and Hands-free Speech Communication and Microphone Arrays (HSCMA 2014).

[14] K. Kinoshita, M. Delcroix, et al., "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013.

[15] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Independent Component Analysis and Signal Separation*. Springer, 2009.

[16] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, 2006.

[17] N. López, M. Maazaoui, et al., "Does dereverberation help multichannel source separation? a case study," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013.

[18] L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 797962, pp. 1–13, 2010.

[19] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, 2003.

[20] E. A. P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[21] L. Wang, H. Ding, and F. Yin, "Speech separation and extraction by combining superdirective beamforming and blind source separation," in *Blind Source Separation*, pp. 323–348. Springer, 2014.