# A NOVEL APPROACH TO MIXED PHASE ROOM IMPULSE RESPONSE INVERSION FOR SPEECH DEREVERBERATION

*Niall Cahill and Robert Lawlor.*

Department of Electronic Engineering,
National University of Ireland, Maynooth,
Maynooth, Co. Kildare,
Ireland.
Email: niall.cahill@eeng.nuim.ie, rlawlor@eeng.nuim.ie.

## ABSTRACT

Outlined in this paper is a novel approach to speech dereverberation when an estimate of the source-receiver transfer function is known. It is a two-stage algorithm based on the minimum phase/allpass decomposition of a mixed phase room impulse response (RIR). The reverberant speech is first filtered with the inverse minimum phase component of the RIR. Then a Non-Negative Matrix Factorization (NMF) based denoising approach is used to remove artifacts associated with the allpass component of the RIR from the inverse filtered speech. This approach was tested on speech convolved with synthetically generated room impulse responses. The results of these tests were analyzed using objective measures and listening tests both of which indicate that this approach leads to significant enhancement of the reverberant speech.

***Index Terms-*** Dereverbation, Non-Negative Matrix Factorization, Room impulse response, Inverse filtering.

## 1. INTRODUCTION

In voice telecommunications, if reverberant speech is transmitted to a distant listener and reproduced it will in general be less intelligible compared with anechoic speech. The speech will often be spectrally colored and sound distant [1]. Dereverberation techniques therefore are useful tools for improving the quality of the speech particularly for hands free telephony.
Reverberent speech *x(n)* is modeled as a convolution of the room impulse response *h(n)* of length *N* and a source speech signal *s(n)*,

$$x(n) = \sum_{m=0}^{N-1} s(m)h(n-m). \qquad (1)$$

The goal of dereverberation is to retrieve the direct path component of the signal. One approach to dereverberation is to obtain a measurement of the impulse response from the source to the receiver and then use the inverse of this measurement to filter the received speech. This approach has a number of problems. Firstly room impulse responses are in general mixed phase responses [2] meaning a straight inversion will result in poles outside the unit circle, leading to either an acausal or unstable inverse filter. Furthermore zeros inside but near the unit circle of the original room response will, when inverted, result in poles that decay very slowly. To accommodate these high Q zeros, inverse filters with a large number of taps would be required. Much work has been devoted to these problems. One early approach [2] was to use homomorphic processing to decompose the RIR into a minimum phase component $h_{mp}$ and an allpass component $h_{ap}$.

$$h(m) = h_{mp}(n) * h_{ap}(n). \qquad (2)$$

Then the guaranteed stable minimum phase component is inverted and is used to filter the speech. For mixed phase responses this magnitude-only equalization is inadequate [2][3] because the remaining allpass component of the mixed phase RIR causes phase distortion in the output speech. This phase distortion has been shown to manifest as audible artifacts in the processed speech [2][3]. In [3] the subjective effects of such processing were studied and a technique for phase equalization was also presented. Mourolpoulas et al [4] proposed another approach to RIR inversion using a linear least square approach with a delay included to compensate for zeros outside the unit circle which was shown to produce better results than minimum phase filtering alone albeit with some artifacts remaining [5]. The linear least squares approximation technique also allows the length of the inverse filter to be controlled so a trade off between filter taps and quality can be introduced.

A challenging aspect of the overall dereverberation problem is when the RIR cannot be measured empirically and can only be estimated from the reverberant speech. In this case before an inverse system can be constructed, the impulse response of the room or the channel must be estimated. This is known as the blind channel estimation problem. A more thorough review of speech dereverberation in general is given in [1].

The focus of this work is on the mixed phase RIR inversion problem; as such we therefore assume knowledge of the RIR. This is also referred to as the acoustic channel inversion problem. It has been observed that when reverberant speech is filtered with the inverse minimum phase component of the inverse filter, distinct audible distortions are remaining. We used a Non-Negative matrix factorization denoising technique presented in [6] to remove these artifacts completely from the speech spectrogram without RIR phase equalization. We show that this approach results in improved speech quality.

This paper is organized as follows: the next section examines the NMF technique with subsections on monaural sound source separation using NMF and a special focus on how it applies to mixed phase RIR inversion. Then the overall methodology is explained in section 3. A description of the experiments performed and results is given in section 4 followed by discussion and conclusions in sections 5 and 6.

## 2. NON-NEGATIVE MATRIX FACTORISATION

Non-Negative Matrix Factorization (NMF) is a linear data analysis technique for non-negative data [7]. The non-negativity constraint of this factorization results in a parts based/additive decomposition of the data where the individual decomposed parts sum together to form the original data. These parts usually capture some structure of the data and provide a more intuitive decomposition [8]. It works by approximating a data set $V \in \mathbb{R}^{\geq 0, M \times N}$ as a multiplication of two matrices $W \in \mathbb{R}^{\geq 0, M \times R}$ and $H \in \mathbb{R}^{\geq 0, R \times N}$.

$$V \approx W \cdot H. \tag{3}$$

The rank of the approximation can be reduced or increased by varying $R$; the number of columns in $W$ and rows in $H$. This usually decreases or increases the reconstruction error depending on the data set. The process of estimating $W$ and $H$ is an optimization problem. Lee and Seung [7] introduced two approaches for estimating $W$ and $H$ each based on a separate cost function. The Euclidean distance between $V$ and $WH$ was one of these cost functions and the second, which was used throughout this work, is a generalized version of the Kullback-Leibler divergence,

$$D(V \| W, H) = \left\| V \odot \log\left(\frac{V}{W \cdot H}\right) - V + W \cdot H \right\|_{Fro}, \tag{4}$$

where $\odot$ is the Hadamard product. The goal of the optimization is to minimize this cost function with respect to $W$ and $H$ whilst imposing the non-negativity constraint. From equation (4) the following multiplicative update rules were derived in [7] to calculate $H$ and $W$,

$$H = H \odot \frac{W^T \cdot \left[\frac{V}{WH}\right]}{W^T \cdot 1}, \quad W = W \odot \frac{\left[\frac{V}{WH}\right] \cdot H^T}{1 \cdot H^T} \tag{5}$$

These update rules are iterated until a prescribed number of iterations has been reached. The updates are alternated between $H$ and $W$, as the objective functions for each are convex separately but not together. Because of the multiplicative updates no update step tuning is needed. The number of iterations specified is data/user dependent and usually picked to occur when cost function $D$ reaches a user-defined threshold.

The matrices $H$ and $W$ will individually express different aspects of the factorization. The columns of $W$ will contain the basis for the data and the rows of $H$ will contain the activation pattern for each basis or the contribution of each basis to the data over time. When multiplied the data is reconstructed with a small error (depending on $R$ and the data).

### 2.1. Convolutive NMF (cNMF) and monaural sound source separation

In [6] Smaragdis presented a cNMF based monaural sound source separation (SSS) technique. The technique requires prior knowledge of the speakers and the order of the mixture. Here we ignore the convolutive extension and concentrate on a one-dimensional manifestation of the algorithm.

The algorithm has two stages, namely training and separation. The training stage involves training separate low rank $W$ basis matrices for each individual speaker. This is done by acquiring a sequence of spoken speech from each speaker, calculating a spectrogram for each sequence and performing a NMF decomposition on each spectrogram separately. The resultant $W$ matrices (one for each speaker) are then concatenated into a large $W$ matrix called $W_{train}$. The second stage is the separation stage where a mixture of speech, containing known speakers, is separated into individual sources. This is achieved by performing a NMF decomposition on the speech mixture using $W_{train}$ from the training stage. Throughout this factorization $W_{train}$ is fixed with only the $H$ matrix updated. This process leads to the basis matrix corresponding to each individual speaker to mainly characterize the mixture spectral energy corresponding to the contribution, which that speaker made to the mixture. After a prescribed number of iterations have been reached $W_{train}$ is separated back to the individual $W$ matrices of the speakers and then multiplied by the corresponding portion of the $H$ matrix from the separation stage. The resultant $V$ matrices are combined with the original phases of the mixture and resynthesised leading to renditions of the original sources.

In [6] the optimal values for these parameters in terms of sound source separation were presented; $R$ (the rank parameter) being particularly important. Apart from algorithm parameters however the best performance was achieved when the mixtures contained one male and one female speaker. This was believed to be due to the level of spectral dissimilarity between the male and female speaker. Between a male and female spectrogram the level of dissimilarity is greater than say between the spectra of two males or two females due to the different pitch tracks and formants etc. As a result of this the trained $W$ matrices for the male and female speech were more easily able to distinguish and better represent their respective contributions in the mixture. This issue of spectral dissimilarity was shown to be an important factor affecting the performance of this algorithm. The algorithm was also tested on noisy speech mixtures and performed well especially for noise that was spectrally dissimilar to speech [6].

### 2.2. Mixed phase room impulse response inversion and NMF

As detailed in section 1, inversion of a mixed phase room impulse response is a difficult task. The inversion problem can however be broken down based on a minimum phase/allpass decomposition of the room impulse response [2]. The minimum phase component can be inverted and used to inverse filter reverberant speech.

The approach results in speech with perceptually annoying artifacts for mixed phase RIRs. These artifacts are a result of the remaining allpass portion of the RIR or more specifically the phase components of the RIR that deviate from a linear group delay [2]. In the literature these artifacts have been described as ‘chimes’ [3] or ‘metallic sounding’ [2]. An explanation for the cause of the artifacts is given in [2] where it is stated that peaks in the allpass group delay function cause the artifacts. In [3] the author describes

the knowledge about the cause of these artifacts as 'still incomplete and conjectural'.

Upon observation of the magnitude spectra of such distorted speech it can be seen that the artifacts are quite distinctive, and appear as smears intermittent throughout the spectrogram (as shown in Figure 1). They also typically occur at similar frequencies; possible the locations of all pass zeros.

The artifacts vary for different RIRs because of the different locations of the allpass zeros. Through repeated visual observation of different spectrograms, the spectral dissimilarity between the artifacts and speech became quite apparent. This motivated us to apply the approach of Smaragdis [6] as a post filtering approach to remove these allpass artifacts from the minimum phase inverse filtered speech.

## 3. TRAINING NMF BASIS FOR ALLPASS ARTIFACT REMOVAL

To use the algorithm described in section 2.1 to remove the allpass artifacts from the inverse minimum phase filtered speech, an initial training stage is required. For this, separate data containing clean speech and artifact data is needed. This separate data will be used as training data from which the $W$ matrices or the bases will learn the distinctive spectral characteristics of the each sound. These two $W$ matrices will then be concatenated to form $W_{train}$ as described in section 2.1.

The training data was acquired in two stages. In the first stage the minimum phase component of the source receiver room impulse response is calculated. This is done using a homomorphic processing technique outlined in [2]. The minimum phase component is then inverted. Then an arbitrary sequence of speech is selected and filtered with the RIR. The reverberant speech is then filtered using the inverse minimum phase component of the RIR leaving the speech and the artifacts. Next the spectrogram of the original speech is subtracted from the processed speech to reveal a spectrum of allpass artifacts. Separate $W$ matrices are then trained for the original speech sequence and the spectrum of allpass components, which are concatenated to form $W_{Train}$. In the denoising or separation stage $W_{Train}$ will be used to extract the artifacts from the speech. It will be shown in the next sections that the training of $W_{train}$ with an arbitrary sequence of speech does not constrain this approach to work on this speech alone.

## 4. EXPERIMENTS

To demonstrate the efficacy of the above-described algorithm a number of experiments were performed and the results evaluated. Two synthetic RIRs were created using the mirror image method of creating room impulse responses [9]. The first RIR (RIR 1) was created in a box room with dimensions 20m (length) × 20 m (width) × 20m (height). A microphone was positioned at the center of the room with a speaker placed 2m in front of the microphone. The second RIR (RIR 2) was simulated in a room of dimensions 8m × 7m × 5m. The microphone for this impulse response was positioned at 4m × 3.5m × 1.2m from the origin of the room and the speaker was placed at 6.6m × 3.5m × 1.2m. The absorption coefficients of the boundaries of each room were varied across frequency to emulate a real room. The approximate $RT_{60}$ of RIR 1 was 1.4 secs and RIR 2 0.3 secs. The RIRs were sampled at 16 kHz and after computation were truncated to 2000 samples (125 ms). The truncation of the RIRs was performed to keep the length of the inverse minimum phase filters computationally reasonable.

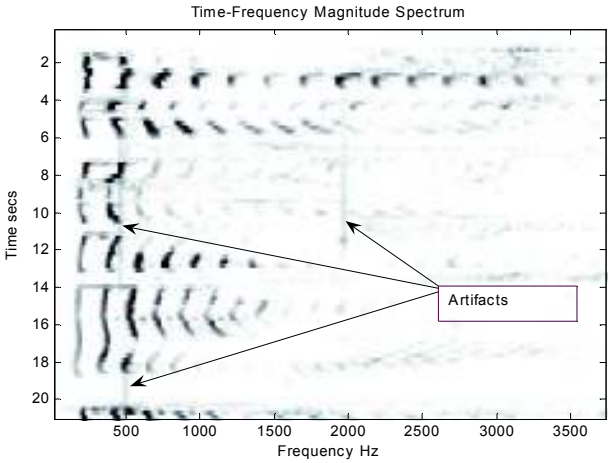


**Figure 1: Time frequency Magnitude spectrum of a Reverberant speech signal after processing by a minimum phase inverse filter with allpass artifacts indicated.**

Moreover to compensate for any high-Q zeros in the truncated RIR, each minimum phase inverse filter was given 20000 taps. The minimum phase component of the RIR was calculated using the approach outlined in [2]. To ensure that the RIRs were mixed phase they were examined using the Nyquist criterion technique also described in [2].

The speech data used for these experiments were taken from the TIMIT speech corpus database [10]. Eight different utterances of speech from 8 different speakers were convolved with each impulse response and were then used for testing. A sequence of speech all from one speaker was chosen arbitrarily as the training speech. The training of $W_{train}$ was performed as described in section 2.1 and 3 with the number of clean speech basis and the number of noise basis set to 100 (i.e. $R$ was set to 100). The number of iterations was also set to 100 for training and testing.

The test speech from the proposed algorithm was analyzed after processing to ascertain the performance of the algorithm. The numerous measures of dereverberation that exist were deemed unsuitable for the approach taken here. To objectively access the output speech a set of standard measures used for sound source separation were employed [11]. These standardized measures include signal to interference noise ratio SIR,

$$SIR = 10\log_{10}\left(\frac{\left\|s_{\text{target}}\right\|^2}{\left\|e_{\text{interf}}\right\|^2}\right), \tag{6}$$

Signal to distortion noise ratio SDR,

$$SDR = 10\log_{10}\left(\frac{\left\|s_{\text{target}}\right\|^2}{\left\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\right\|^2}\right), \tag{7}$$

and signal to artifact noise ratio SAR,

$$SAR = 10\log_{10}\left(\frac{\left\|s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\right\|^2}{\left\|e_{\text{artif}}\right\|^2}\right). \tag{8}$$

Where $s_{target}$ is the output speech, $e_{artif}$ is the processing artifacts noise, $e_{noise}$ processing noise and $e_{interf}$ the interfering noise (i.e. the allpass artifacts). These measures were introduced in [11] to provide a standardized approach to evaluate sound source separation algorithms. To utilize these measures here we consider

(a)

| RIR 1 | Input SIR | Input SDR | Input SAR | Output SIR | Output SDR | Output SAR |
|---|---|---|---|---|---|---|
| | 15.06 | 11.90 | 14.90 | 26.67 | 9.77 | 9.87 |
| | 25.28 | 21.45 | 23.79 | 47.83 | 14.61 | 14.62 |
| | 20.28 | 17.17 | 20.12 | 36.64 | 15.54 | 15.57 |
| | 24.54 | 20.28 | 22.34 | 39.80 | 12.56 | 12.57 |
| | 20.34 | 17.86 | 21.52 | 31.99 | 10.81 | 10.84 |
| | 17.15 | 12.91 | 15.04 | 37.87 | 10.46 | 10.47 |
| | 23.45 | 19.80 | 22.27 | 65.72 | 10.27 | 10.27 |
| | 20.17 | 17.93 | 21.92 | 31.55 | 7.56 | 7.59 |
| **mean** | 20.71 | 17.20 | 20.12 | 39.13 | 12.07 | 12.10 |
| **RIR 2** | 12.99 | 14.59 | 10.62 | 20.02 | 11.47 | 12.17 |
| | 14.00 | 12.56 | 10.12 | 26.57 | 8.96 | 9.04 |
| | 10.85 | 9.55 | 6.94 | 20.87 | 7.20 | 7.43 |
| | 10.86 | 10.93 | 7.71 | 19.85 | 8.36 | 8.72 |
| | 14.48 | 11.41 | 9.56 | 22.98 | 9.00 | 9.19 |
| | 10.49 | 13.50 | 8.60 | 16.75 | 8.04 | 8.76 |
| | 10.48 | 13.17 | 8.48 | 17.51 | 9.91 | 10.82 |
| | 7.59 | 10.15 | 5.41 | 16.34 | 4.83 | 5.25 |
| **mean** | 11.47 | 11.98 | 8.43 | 20.11 | 8.47 | 8.92 |

(b)

| Tag | prefer | >prefer |
|---|---|---|
| FCJF0 | 7 | 3 |
| MDPK0 | 5 | 5 |
| FETB0 | 6 | 4 |
| MJWT0 | 6 | 4 |
| FSAH0 | 8 | 2 |
| MWAD0 | 7 | 3 |
| FVFB0 | 7 | 3 |
| MRWS0 | 4 | 6 |
| | 6.25 | 3.75 |
| FCJF0 | 6 | 4 |
| MDPK0 | 5 | 5 |
| FETB0 | 7 | 3 |
| MJWT0 | 5 | 5 |
| FSAH0 | 7 | 3 |
| MWAD0 | 7 | 3 |
| FVFB0 | 6 | 4 |
| MRWS0 | 4 | 6 |
| | 5.875 | 4.125 |

**Table 1: (a) Table of objective results for RIR 1 and RIR 2. (b) Subjective results for listening tests with TIMIT speech tags and for RIR1 and RIR2.**

the artifacts and the clean speech as two separate sound sources and calculate the objective measures based on this.

To subjectively access the performance of the algorithm an informal listening test was organized. Ten subjects were recruited and asked to compare two utterances of the same speech, one was the speech with artifacts (minimum phase inverse filtered) and the other was the speech processed using the new approach. They were then asked to give their opinion based on five ratings: much prefer A, prefer A, neutral, prefer B, much prefer B. The order of the two utterances was randomized to mitigate bias and each sentence pair was compared for RIR 1 and 2. The number of preferences given by the panel for each rating is listed in table 1 (b). Audio examples of the above work are available at www.eeng.nuim.ie/~ncahill/.

## 5. DISCUSSION

The allpass artifacts remaining in the speech after filtering by the minimum phase inverse filter were removed or significantly suppressed by this new algorithm. The results of the listening tests in table 1 (b) show that the subjects overwhelming preferred the post processed sound. Specifically no subjects choose neutral or the opposing utterance and a large amount of subjects rated the processed speech as 'much preferred' or '> prefer' in table 1(b). The objective measures in table 1(a) also demonstrate the improvement of this approach. Large gains in SIR suggest the post filtering stage greatly removed the artifacts whilst a lower SAR and SDR indicate that the algorithm also removed some clean speech. The experimental results also demonstrate that training based on one speaker does not restrict the algorithm to this speaker or to the utterances used. However the above approach is impulse response dependent and each $W_{train}$ is only usable for the RIR used for training. Moreover these results indirectly show that the phase distortion introduced by the allpass component of mixed phase RIRs can be greatly suppressed by post-filtering the magnitude spectrum of such distorted speech.

## 6. CONCLUSIONS

In this paper a new approach to acoustic channel inversion of mixed phase room impulse responses was introduced. The process starts with a conventional minimum phase/allpass decomposition followed by inverse minimum phase processing. The new approach then uses a NMF based post filtering stage to remove the allpass artifacts. The results of experiments designed to test the algorithm were presented which show the improvement in speech quality.

## 11. REFERENCES


[1] Naylor P.A. and Gaubitch N.D., "Speech Dereverberation" Proc. *International Workshop on Acoustic Echo and Noise Control (IWAENC 2005)*

[2] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 165–169, July 1979.

[3] B. Radlovic, R. Kennedy, "Nonminimum-Phase Equalization and Its Subjective Importance in Room Acoustics," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, Nov 2000

[4] Mourjopoulos J., "Digital Equalization of Room Acoustics", *Journal of the Audio Engineering Society,* Vol.42, No.11, pp. 884- 900, Nov. 1994.

[5] Mourjopoulos J.N., "Comments on "Analysis of Traditional and Reverberation-Reducing Methods of Room Equalization", *Journal of the Audio Engineering Society,* Vol. 51, No.12, pp. 1186-1188, December 2003

[6] Smaragdis, P., "Convolutive Speech Bases and their Application to Supervised Speech Separation", *IEEE Transaction on Audio, Speech and Language Processing,* Vol. 15, Issue 1, pp. 1-12, January 2007

[7] Lee, D.D., Seung, H.S. "Algorithms for non-negative matrix factorization", in *Advances in Neural Information Processing Systems* 13, 2000.

[8] Lee, D.D., Seung, H.S. "Learning the Parts of Objects by Nonnegative Matrix Factorization", in *Nature* 1999 (401):788.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr.1979.

[10] Fisher W.M., Doddington G.R., and Goudie-Marshall K, "The DARPA Speech Recognition Research Database: Specifications and Status," *Proceedings of DARPA Workshop on Speech Recognition,* pp. 93-99, Feb. 1986.

[11] Vincent E.; Gribonval R.;Fevotte C.; "Performance Measurement in Blind Audio Source Separation", *IEEE trans on Speech and Audio processing.* Volume PP, Issue 99, 2005 Page(s): 1 – 8.