A NMF based Joint Speech Dereverberation and Denoising Utilizing Different RIR Spectrogram Structures

A Thesis
Submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
by

Nikhil M
(Roll No. 144076002)

Supervisors:
Prof. Rajbabu V
and
Prof. Preeti Rao

Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 400076 (India)

5 May 2020

Dedicated to . . .

Acceptance Certificate

Department of Electrical Engineering
Indian Institute of Technology, Bombay

The thesis entitled "A NMF based Joint Speech Dereverberation and Denoising Utilizing Different RIR Spectrogram Structures" submitted by Nikhil M (Roll No. 144076002) may be accepted for being evaluated.

Date: 5 May 2020

Prof. Rajbabu V

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I declare that I have properly and accurately acknowledged all sources used in the production of this report. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Nikhil M

Date: 5 May 2020 (Roll No. 144076002)

Abstract

Speech recorded in a distant speech recording (DSR) scenario is corrupted by reverberation and noise resulting in poor quality recorded speech. This work proposes a novel non-negative matrix factorization (NMF) based single-channel enhancement method to handle reverberation and noise jointly. Such an approach is different from other NMF based approaches in the literature, which use a combination of convolutive NMF (CNMF) and NMF to model reverberation and noise. The proposed NMF model introduces better constraints on the room impulse response (RIR) that is not possible using other NMF based approaches. We achieve the proposed NMF representation by introducing a low-rank factorization for the magnitude spectrogram of the RIR. We show that such a form accurately represents the RIR spectrogram. Further, based on the RIR model, a supervised joint dereverberation and denoising method is proposed. The enhancement method is extended to work in unknown noise conditions. The performance of the proposed method has been validated on degraded utterances simulated from the TIMIT dataset and compared with other NMF based enhancement approaches. The objective measures indicate that the proposed enhancement method performs consistently better than other methods. The proposed method also gives a better estimate of the RIR magnitude spectrogram.

Table of Contents

Abstract **ix**

List of Figures **xv**

List of Tables **xvii**

List of Symbols **xix**

1 Introduction **1**

1.1 Motivation 1

1.2 Distant Speech Recording (DSR) 1

1.3 Reverberation 1

1.4 Objective of the thesis 3

1.5 Contributions 3

1.6 Organization of thesis 3

2 Related Works **5**

2.1 Dereverberation methods 5

2.1.1 Beamforming 5

2.1.2 Inverse filtering based methods 7

2.1.3 Reverberation suppression methods 9

2.2 NMF based dereverberation methods 9

2.3 Limitations of NMF based dereverberation methods in literature 12

3 Properties of RIR **13**

3.1 Structure of RIR 13

3.1.1 Time domain structure 13

3.1.2 Magnitude spectrum of RIR 14

3.1.3 Pole-zero plot of RIR 15

3.1.4 Inverse filtering 15

3.2 Magnitude spectrogram of RIR . . . 17
3.2.1 Pollack's model . . . 17
3.2.2 Spectrogram structure of RIR . . . 17
3.3 Characteristics of RIR . . . 22
3.3.1 Source-microphone distance (d_{sd}) . . . 22
3.3.2 Reverberation time (T_{60}) . . . 23
3.3.3 Direct-to-reverberation ratio (DRR) . . . 23
3.4 Conclusion . . . 24

4 Dereverberation based on RIR constrained cost function 25
4.1 Basic NMF based dereverberation methods . . . 25
4.1.1 CNMF based reverberation model . . . 25
4.1.2 CNMF model with NMF model for clean speech . . . 27
4.1.3 Comparison of reverberation models . . . 29
4.2 Incorporation of RIR properties on optimization problem . . . 29
4.2.1 Frequency envelope of RIR spectrogram . . . 30
4.2.2 Sparsity of RIR spectrogram . . . 30
4.2.3 Retaining early part of RIR . . . 30

5 Separability Assumption on RIR Spectrogram 31
5.1 NMF degradation models . . . 31
5.2 Analysis of model . . . 31
5.3 Enhancement algorithm . . . 31
5.4 Experimental results . . . 31
5.5 Discussion . . . 32

6 Low-rank NMF model for RIR spectrogram 33
6.1 Justification of low-rank approximation . . . 33
6.2 NMF model for degraded spectrogram . . . 33
6.3 Algorithm details . . . 33
6.4 Experimental results . . . 34
6.5 Discussion . . . 34

7 Conclusion and suggestion for future work 33

A Supporting Material 35

References 37

List of Publications **41**

Acknowledgements **43**

List of Figures

1.1 DSR scenario 2

1.2 Thesis chapters 4

2.1 Two-microphone DSR recording setup. Microphones are placed d-distant apart. The effects of reverberation and noise is not shown in the figure. . . 6

3.1 The magnitude spectrogram (left) and temporal envelope for different frequency bands (right) for a measured RIR with $T_{60} \approx 700$ ms and source-microphone distance of 2 m. The temporal envelope decays down with time as is the case in Pollack's model. 19

3.2 (a) Frequency envelope and temporal variation obtained for a measured RIR from [1] with $T_{60} \approx 700$ ms and source-to-microphone distance d = 2 m. (b) Frequency envelope and temporal variation obtained by a rank-1 NMF decomposition of the RIR. Frequency envelope approximated in (b) captures the most variations in (a). 20

3.3 Effect of varying rank P on the low-rank approximation for the RIR spectrogram. The deviation from the original RIR spectrogram reduces with increasing P. The deviation is small for $P > 10$. 22

3.4 (a) Frequency envelope and temporal variation obtained for a measured RIR from [1]. (b) Frequency envelope and temporal variation obtained by a rank-10 NMF decomposition of the RIR. (b) is a very good approximation of (a). 23

List of Tables

6.1 Enhancement results when reverberated with RIR RIR3_far and stationary noise added with 10 dB SNR. The RIR has $T_{60} \approx 700$ ms and source-microphone distance of 2 m. 34
6.2 Enhancement results for 10 dB SNR stationary noise. 34
6.3 Enhancement results for 20 dB SNR stationary noise. 35
6.4 Enhancement results for noise free condition. 35

List of Symbols

Roman Symbols

R Radius of circle 4

r Intrinsic length 4

Greek Symbols

θ Incidence angle 4

Superscripts

g Gas phase 4

v Vapor phase 4

Subscripts

R Reverberation 4

Acronyms

DSR Distant speech recording 4

RIR Room impulse response 4

ASR Automatic speech recognition 4

TDOA Time difference of arrival 4

LS Least square method 4

Other Symbols

$s(n)$ Time-domain clean speech signal 4

$h(n)$	Time-domain RIR	4
$y_R(n)$	Time-domain reverberated speech signal	4
$y(n)$	Time-domain reverberated and noisy speech signal	4
$z(n)$	Time-domain noise	4
d	Microphone spacing	4
v	Velocity of sound	4
$y_{R,m}(n)$	Microphone output for m-th element in a microphone array in reverberant condition	4
$h_m(n)$	RIR form the source to m-th element in a microphone array	4
$\mathbf{h}_m$	Vector whose elements are obtained from the m-th element in a microphone array $h_m(n)$	4
$\mathbf{h}$	Vector obtained from concatenated RIRs in a microphone array	4
$\mathbf{R}$	Correlation-like matrix obtained from clean speech $s(n)$	4

Chapter 2

Related Works

This chapter discusses various speech enhancement methods available in the literature. The different class of dereverberation methods available in the literature is summarized in Section 2.1. Since this work is based on NMF based model for reverberation and noise, the NMF based enhancement methods in the literature are explained in Section 2.2.

2.1 Dereverberation methods

This section discusses various dereverberation methods in literature. Many of these methods can be extended to hande reverberation in presence of noise. The dereverberation methods can be broadly classified into - (i) beamforming, (ii) inverse filtering based methods, and (iii) reverberation suppression methods. Each methods are discussed in details next.

2.1.1 Beamforming

Beamforming is a multi-channel method. It utilizes the spatial information of the source and microphone array to enhance degraded speech recordings. A beamformer enhances the signal received from a particular direction and attenuates the signal received from other directions [2]. This spatial filtering is made possible by the fact that the sound waves travel an additional distance to reach distant microphones when compared with nearer microphones. This result in a relative time lag is referred to as time delay of arrival (TDAO). TDOA depends on the source position and microphone array configuration.

Figure 2.1 illustrates the occurrence of TDOA for a source paced distant from a two-microphone array. The source is placed at an angle θ from the axis of the microphone array (referred to as incident angle). The signal travels an extra distance of $d\cos(\theta)$ to reach microphone M_1 when compared with the reference microphone M_{ref}. This results

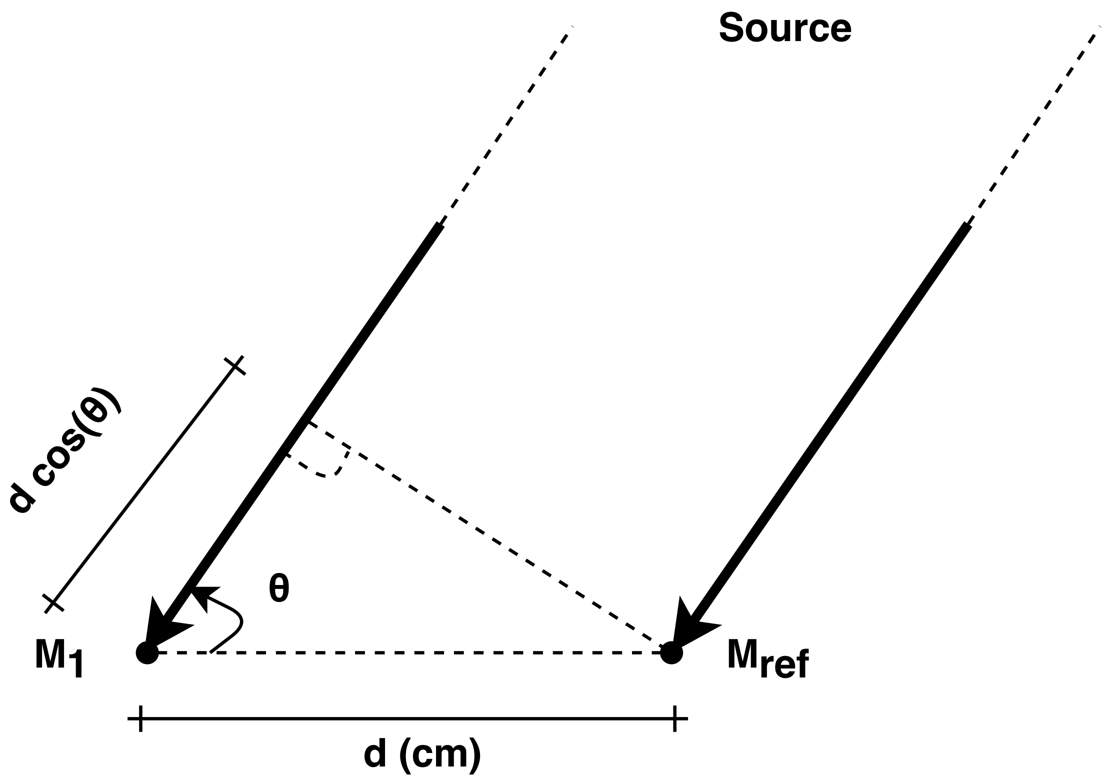

Figure 2.1: Two-microphone DSR recording setup. Microphones are placed d-distant apart. The effects of reverberation and noise is not shown in the figure.

in a time lag τ of

$$\tau = \frac{d\cos(\theta)}{v}, \tag{2.1}$$

where v is velocity of sound in the medium. It can be observed that TDOA changed with the incident angle θ and microphone spacing d.

Delay-sum beamforming (DSB) is the most straight forward beamforming approach. The microphone recordings are delayed to compensate for different TDOAs and a convex combination of these signals is taken. For a M-microphone array, the enhanced signal $\hat{x}(n)$ obtained using DSB can be summarized as,

$$\hat{x}(n) = \sum_{m=1}^{M} p_m x_m(n - \tau_m), \tag{2.2}$$

where $x_m(n)$ is the recording for m-th microphone. p_m and τ_m represents the weight and TDOAs obtained for each microphone. p_m can be fixed based on some amplitude normalization criteria [5]. τ_m can be estimated based on the source localization algorithms like generalized cross-correlation with phase transform (GCC-PHAT) [6]. The signals from look-direction in different channels will add constructively. This results in enhancing signals coming from look-direction at the expense of the signals received from other directions. Beamforming was shown to effectively suppress localized noise. However, reverberant speech is partially suppressed. Reverberation makes localized speech sources into diffused sound. The resulting reverberated speech signal reaches the array from all directions. Hence, reverberated speech signal coming from look direction will not be suppressed, while from other directions will be suppressed.

Many modifications were proposed to improve the performance of the beamformer. MVDR beamforming uses noise statistics to improve performance. The effects of reverberation were suppressed with the use of multiple beamformers. This is achieved with the use of a three-dimensional microphone array. One beam is steered in the direction of the desired location as was the case in earlier. Additional beams are steered in the direction of the strong initial reflections [7] which acts as virtual sources. Another method to suppress reverberation was the use of a matched filter beamformer where the microphone responses are convolved with time-reversed RIR. The method requires Reverberation suppression methods model the effect of reverberation on the clean speech in some domains like the short-time Fourier transform (STFT), the residual signal obtained from linear prediction analysis, etc. Based on these models, algorithms are proposed to reduce the effects of reverberation. This results in a better estimate of clean speech. The approaches include spectral subtraction [19], weighted linear prediction (WPE) [20], Triple-N ICA for convolutive mixtures (TRINICON) [21], linear prediction based method [2], and NMF based approaches [22], information about RIR [2].

2.1.2 Inverse filtering based methods

Inverse filtering based methods are multi-channel speech enhancement methods [2]. These methods have two steps - (i) blind estimation of RIR and (ii) inverse filtering.

(i) Blind estimation of RIR

In this step, a blind estimate for RIR from the source to at least one of the microphones in the array is obtained. The correlation between speech recorded by different microphone channels is used for the purpose. Consider two channels DSR recording case degraded by reverberation alone. The microphone outputs $y_{R,1}$ and $y_{R,2}$ represented as,

$$\begin{aligned} y_{R,1}(n) &= s(n) * h_1(n) \text{ and} \\ y_{R,2}(n) &= s(n) * h_2(n) \end{aligned} \tag{2.3}$$

where $s(n)$ is the clean speech, and $h_1(n)$ and $h_2(n)$ represents the RIRs form the source to the first and the second microphone, respectively. The microphone outputs are correlated based on the following relation.

$$\begin{aligned} y_{R,1}(n) * h_2(n) &= (s(n) * h_1(n)) * h_2(n) \\ &= h_1(n) * (s(n) * h_2(n)) \\ &= y_{R,2}(n) * h_1(n) \\ y_{R,1}(n) * h_2(n) - y_{R,2}(n) * h_1(n) &= 0 \end{aligned} \tag{2.4}$$

Based on (2.4), a system of equations of the following form can be written [8].

$$\mathbf{Rh} = \mathbf{0} \tag{2.5}$$

where, $\mathbf{R}$ represents a correlation-like matrix obtained from source signal $s(n)$ [9]. $\mathbf{0}$ represents a zero vector. The vector $\mathbf{h}$ is obtained from RIRs as $\mathbf{h} = [\mathbf{h}_1^T|\mathbf{h}_2^T]^T$, where $h_1(n)$ and $h_2(n)$ form the elements of $\mathbf{h}_1$ and $\mathbf{h}_2$, respectively. The ideal solution for the $\mathbf{h}$ in the system of equations (2.5) is the eigen-vector corresponding to the zeroth eigen-value in $\mathbf{R}$. In the presence of noise, the solution for $\mathbf{h}$ is eigen-vector corresponding to the smallest eigen-value. The system of equations in (2.5) can be generalized for more microphone recordings are available.

Different approaches have been proposed in the literature to blindly estimate RIR based on the system of equations in (2.5). Such methods assume that certain identifiability conditions and practical considerations are met [2]. Some of these methods are explained here. In [9], a least-squares (LS) approach was proposed to solve the system of equations in (2.5). The method based on eigendecomposition was also been proposed [10]. In [11], full bands and frequency sub-bands eigendecomposition methods were proposed to solve for (2.5) in the presence of white and colored noises. Adaptive filter in time domain [12] and frequency domain [9] were also proposed in literature.

(ii) Inverse filtering

Inverse filtering methods could be used to estimate clean speech $s(n)$ from reverberated microphone recordings $y_{R,m}(n)$ and an estimate of RIRs $h_m(n)$. The most straight forward method would be to use an inverse filter $\mathbf{G}_m$ that compensate for effects of reverberation as shown next.

$$\mathbf{y}_m^T\mathbf{G}_m = \alpha_0\delta(n - n_0), \tag{2.6}$$

where α_0 and n_0 represent arbitrary scaling and delay factors, respectively. $\delta(n)$ represents unit discrete delta function. However, the use of the direct inverse system shown in (2.6) is challenging because of the following factors. (i) RIR has non-minimum phase characteristics [13], (ii) spectral nulls can be present RIR spectrum and (iii) estimated inverse filter will have thousands of coefficients that require high precision and computationally expensive methods. In [14], inverse filter estimate $\hat{\mathbf{G}}_m$ was obtained based on a LS method. The estimated inverse filter minimizes the following cost function.

$$\hat{\mathbf{G}}_m = \min_{\mathbf{G}_m} \left\| \mathbf{h}_m^T \mathbf{G}_m - \delta(n - n_0) \right\|_2^2 \tag{2.7}$$

Homomorphic inverse filtering methods were also been investigated [15, 14]. The inverse filter was decomposed into minimum-phase and all-pass components. The minimum-phase component can be directly estimated from the magnitude spectrum of estimated RIR. Various methods like matched filtering [15] were used to estimate the all-pass components.

When multi-channel RIRs are available, multiple-input/output inverse theorem (MINT) based approaches can be used to find exact inverse filters [16]. According to the theorem, if the transfer function of two RIRs does not have any common zeros, then there exist a pair of inverse filters $\mathbf{G}_1$ and $\mathbf{G}_2$ such that

$$\mathbf{h}_1^T \mathbf{G}_1 + \mathbf{h}_2^T \mathbf{G}_2 = \delta(n) \tag{2.8}$$

In [16], a LS method was used to solve for (2.8). An exact inverse filtering can be performed by using an inverse filter length similar to that of RIR length. A sub-band version was proposed in [17]. An adaptive version of the method is proposed in [18]. Regularization was imposed on equalization problem in (2.8) to improve the robustness against noise and estimation errors [2].

2.1.3 Reverberation suppression methods

Reverberation suppression methods are based on modeling the effect of reverberation on clean speech in some domains. The commonly used domains are the short-time Fourier transform (STFT) and its variants, the residual signal obtained from linear prediction analysis, etc. Based on these degradation models, algorithms were proposed to reduce the effects of reverberation. This results in a better estimate of clean speech. The approaches include spectral subtraction [19], weighted linear prediction (WPE) [20], Triple-N ICA for convolutive mixtures (TRINICON) [21], linear prediction based method [2], and NMF based approaches [22], etc. SuchTypically these algorithms are single-channel methods. However, these methods can easily be extended for a multi-channel scenario.

2.2 NMF based dereverberation methods

NMF based speech enhancement methods are based on the modulation transfer function (MTF) model for reverberation. The MTF model states that for each frequency sub-band, the power envelope of the reverberated speech $\mathbb{Y}_R(k,n)$ is the convolution of power envelopes of clean speech $\mathbb{S}(k,n)$ and RIR $\mathbb{H}(k,n)$ for that particular sub-band. Mathematically,

$$\mathbb{Y}_R(k,n) = \mathbb{H}(k,n) *_n \mathbb{S}(k,n) = \sum_{l=0}^{L_h-1} \mathbb{H}(k,l)\mathbb{S}(k,n-l), \tag{2.9}$$

where $*_n$ represents convolution along the time axis. L_h represents the number of frames used to represent the RIR spectrogram. This model is valid when the reverberation condition does not change with time. Further, the MTF model is obtained by ignoring the cross-band effects occurring due to windowing [23]. The RIR phase spectrogram is also assumed to be uniformly distributed in the range $[-\pi, \pi)$. Even though these approximations in the MTF model can pose a limitation for the dereverberation task, many dereverberation methods use this model. An additional advantage of the MTF model is that it avoids the need for a phase estimate for the RIR. Obtaining the phase spectrogram is difficult, especially if the recording is noisy [22]. Estimation of $\mathbb{H}(k,n)$ and $\mathbb{S}(k,n)$ from $\mathbb{Y}_R(k,n)$ in (2.9) is viewed as solving for a CNMF problem. The dereverberated speech obtained using this algorithm showed improvements in speech enhancement instrumental measures.

Many modifications were proposed to improve the performance. The use of a magnitude spectrogram instead of a power spectrogram showed superior performance [24]. The magnitude spectrogram of reverberated speech $\mathbf{Y}_R$ is expressed as a convolution of magnitude spectrograms of clean speech $\mathbf{S}$ and RIR $\mathbf{H}$. Mathematically,

$$\begin{aligned} \mathbf{Y}_R &= \mathbf{H} *_n \mathbf{S} \\ Y_R(k,n) &= \sum_{l=0}^{L_h-1} H(k,l)S(k,n-l), \end{aligned} \tag{2.10}$$

where, $Y_R(k,n)$ $H(k,n)$, and $S(k,n)$ represents the elements of $\mathbf{Y}_R$, $\mathbf{H}$ and $\mathbf{S}$, respectively. Further, it was shown that the use of gamma-tone filter banks helped in improving ASR results when compared with the use of uniform filter banks.

The reverberation model in (2.10) did not use any model for the clean speech spectrogram. Introducing clean speech spectrogram models were shown to improve speech enhancement results. In [25, 26], a NMF model for the clean speech spectrogram $\mathbf{S}$ was incorporated in the reverberation model in (2.10). The NMF model of clean speech spectrogram can be written as,

$$\begin{aligned} \mathbf{S} &= \mathbf{W}_s\mathbf{X}_s \\ S(k,n) &= \sum_{r=1}^{R_s} W_s(k,r)X_s(r,n), \end{aligned} \tag{2.11}$$

where, $\mathbf{W}_s$ and $\mathbf{X}_s$ represents bases and activation matrices obtained by performing NMF decomposition on the clean speech spectrogram $\mathbf{S}$, respectively. R_s represents the rank of NMF decomposition. $W_s(k, r)$ and $X_s(r, n)$ represents the elements of $\mathbf{W}_s$ and $\mathbf{X}_s$, respectively. Incorporating the clean speech model (2.11) in the reverberation model in (2.10) results in a reverberation model that can be written as,

$$Y_R(k, n) = \sum_{l=0}^{L_h-1} H(k, l)\Big[\sum_{r=1}^{R_s} W_s(k, r)X_s(r, n - l)\Big] \tag{2.12}$$

Iterative algorithm were proposed to estimate clean speech spectrogram. There were two approaches depending on how clean speech bases $\mathbf{W}_s$ is learned - online approach and offline approach. In offline approach , the $\mathbf{W}_s$ is learned from reverberated spectrogram. In online approach, $\mathbf{W}_s$ is pre-learned from a set of clean speech utterances. A NMF decomposition is performed on the magnitude spectrogram of these clean speech utterances to estimate $\mathbf{W}_s$. Many constraints on the estimated clean speech like sparsity [25, 26], and continuity [27] were used to improve the performance.

Representing clean speech spectrogram using CNMF model were also proposed to improve speech dereverberation perfromance [28]. In [29], a NMF model for speech reverberation was proposed. This model is equivalent to the CNMF model for reverberation in (2.12). The bases matrix of the NMF decomposition is a structured matrix that was constructed to mimics the CNMF model.

The reverberation model in (2.12) is inappropriate to model reverberation in the presence of noise. The model was modified by incorporating a noise model. The magnitude spectrogram of reverberated speech in the presence of noise $\mathbf{Y}$ was approximated as the sum of magnitude spectrograms of reverberated speech $\mathbf{Y}_R$ and noise $\mathbf{Z}$ [30]. Mathematically,

$$\begin{aligned} \mathbf{Y} \approx \mathbf{Y}_R + \mathbf{Z} = \mathbf{H} *_n \mathbf{S} + \mathbf{Z} \\ Y(k, n) = H(k, n) *_n S(k, n) + Z(k, n), \end{aligned} \tag{2.13}$$

where $Y(k, n)$ and $Z(k, n)$ represents the elements of $\mathbf{Y}$ and $\mathbf{Z}$, respectively. Similar to the NMF model for clean speech spectrogram in (2.11), a NMF approximation for noise spectrogram can be used as shown in (2.14).

$$\begin{aligned} \mathbf{Z} = \mathbf{W}_n\mathbf{X}_n \\ Z(k, n) = \sum_{r=1}^{R_n} W_n(k, r)X_n(r, n), \end{aligned} \tag{2.14}$$

where $\mathbf{W}_n$ and $\mathbf{X}_n$ were the bases and activation matrix of noise spectrogram $\mathbf{Z}$. R_n represents the rank of NMF decomposition. $W_n(k, r)$ and $X_n(r, n)$ represents the elements of

$\mathbf{W}_n$ and $\mathbf{X}_n$, respectively. The degradation model in (2.13) is modified with the use of NMF models for clean speech in (2.11) and noise (2.14) as shown next.

$$\mathbf{Y} = \mathbf{H} *_n \left[\mathbf{W}_s\mathbf{X}_s\right] + \mathbf{W}_n\mathbf{X}_n \tag{2.15}$$

Based on the degradation model in (2.15), a speech enhancement algorithm was proposed in [31]. The algorithm was a supervised approach where the clean speech and noise bases are assumed to be known. Exemplar-bases are learned for the purpose. This approach of obtaining the bases is different from the one used in [25, 26]. The use of the CNMF model for the clean speech and the noise spectrogram was also proposed in [31, 32]. The use of the CNMF model for the clean speech and the noise spectrogram was also proposed in [31, 32]. In [31], the temporal variation of the RIR spectrogram was modeled along with the degradation model in (2.15).

2.3 Limitations of NMF based dereverberation methods in literature

The NMF based speech enhancement methods in literature utilize limited information about the RIR spectrogram. This limits the performance of these algorithms. In this work different spectro-temporal models for the RIR spectrograms are proposed. Incorporating these novel RIR constraints on the NMF based enhancement algorithm helped in improving the performance. The reverberation models used in this work models the effects of reverberation on the magnitude spectrogram of clean speech. Hence, in the subsequent chapters, the usage of spectrogram means the magnitude spectrogram, unless stated otherwise. Chapter 3 discusses the various properties of the RIR. Some of these properties are used in this work. The subsequent chapters discuss the different proposed speech enhancement algorithms that utilize these RIR models.

Chapter 4

Dereverberation based on RIR constrained cost function

This chapter explains NMF based dereverberation methods that utilize three different properties of RIR spectrogram - sparsity, frequency envelope, and early part of RIR. Such methods are derived by modifying the basic NMF based dereverberation methods available in the literature. The initial section of this chapter explains the basic NMF based dereverberation problems. Later part of the chapter explains the modifications made to accommodate the above mentioned RIR properties in the dereverberation problem. This discussion is followed by the analysis of enhancement results.

4.1 Basic NMF based dereverberation methods

This section explains the approach taken to estimate clean speech and RIR spectrogram based on the basic reverberation models discussed in Section 2.2. The algorithms based on the CNMF reverberation model in (2.10) and CNMF reverberation model with NMF clean speech model in (2.12) is discussed next.

4.1.1 CNMF based reverberation model

The CNMF model for reverberated speech spectrogram was discussed in Section 2.2. In (2.10), the reverberated speech spectrogram $Y_R \in \mathbb{R}_+^{K\times T}$ is approximated as convolution of clean speech spectrogram $\mathbf{S} \in \mathbb{R}_+^{K\times(T-L_h+1)}$ and RIR spectrogram $\mathbf{H} \in \mathbb{R}_+^{K\times L_h}$.

$$Y_R(k,n) \approx \sum_{l=0}^{L_h-1} H(k,l)S(k,n-l), \tag{4.1}$$

where $Y_R(k,n)$, $H(k,n)$ and $S(k,n)$ represents the elements of $\mathbf{Y_R}$, $\mathbf{H}$ and $\mathbf{S}$, respectively. Iterating algorithm was proposed for solving for $S(k,n)$ and $H(k,n)$ based on reverberation model in (4.1) [22, 24]. The parameters are estimated such that it minimizes a cost function. Euclidean distance (ED) [22] and generalized KL divergence [24] are commonly used cost functions. Generalized KL divergence (KL) as cost functions gives reduced modeling error when compared to ED. The optimization problem for solving the CNMF based reverberation model based on KL divergence can be written as,

$$C_{cnmf0} = \underset{S(k,n),\, H(k,n)}{\operatorname{argmin}} \left[\sum_{k,n} \mathrm{KL}\Big(Y_R(k,n) | \tilde{Y}_R(k,n)\Big) + \lambda_{cnmf0} \sum_{k,n} S(k,n) \right]$$

Subjected to

$$\begin{aligned} & H(k,n) \geq 0,\, S(k,n) \geq 0 \\ & \sum_{n} H(k,n) = 1, \end{aligned} \tag{4.2}$$

where $\tilde{Y}_R(k,n)$ represents the estimated reverberated spectrogram based on (4.1). λ_{cnmf0} is a weighting factor. The cost function in (4.2) has two terms. First term is a measure of deviation between actual and estimated reverberated speech spectrogram. The second term introduces sparsity to the estimated clean speech spectogram. The amount of sparsity is controled by λ_{cnmf0}. It is fixed as $\lambda_{cnmf0} = \frac{10^{-8}}{KT} \sum_{k,n} Y_R(k,n)$. This first set of constraints make sure that the estimated $S(k,n)$ and $H(k,n)$ are non-negative. This is necessary as these terms represents magnitude spectrograms. The normalization $\sum_{n} H(k,n) = 1$ was required to avoid scaling ambiguity and the subsequent estimation of undesired solution [1]. Iterative algorithm based on a multiplicative update rule was proposed to find the solution [24]. The update rule for the parameters are shown in (4.3). This method is referred to as CNMF0.

$$\begin{aligned} H(k,n) &\leftarrow H(k,n) \frac{\sum_{l} \frac{Y_R(k,l)}{\tilde{Y}_R(k,l)} S(k,n-l)}{\sum_{l} S(k,n-l)} \\ S(k,n) &\leftarrow S(k,n) \frac{\sum_{l} \frac{Y_R(k,l)}{\tilde{Y}_R(k,l)} H(k,n-l)}{\sum_{l} H(k,n-l) + \lambda_{cnmf0}} \end{aligned} \tag{4.3}$$

The steps involved in CNMF0 is summarized in Algorithm 1. The time-domain de-reverberated speech is obtained by performing an inverse STFT on the complex spectrogram of the estimated clean speech. The complex spectrogram is constructed by using

[1]Assume a clean speech spectrogram $S(k,n)$ and RIR $H(k,n)$ minimizes the first term in (4.1). Then $aS(k,n)$, $a \in \mathbb{R}_+$ and $H(k,n)/a$ will also minimizes the first term. So, the cost function (4.1) is minimized by the value that minimizes the second term. This happens when $S(k,n) = 0$, resulting in $H(k,n) = \infty$.

the phase spectrogram obtained from the reverberated speech along with the enhanced magnitude spectrogram.

Algorithm 1: Steps involved in CNMF0

Result: Enhanced speech spectrogram $S(k,n)$

initialize $S(k,n)$, $H(k,n)$ to random positive values;

for $i = 1 : i\_max$ **do**

  update $S(k,n)$

  update $H(k,n)$

  normalization $H(k,n) \leftarrow \dfrac{H(k,n)}{\sum_n H(k,n)}$

end

4.1.2 CNMF model with NMF model for clean speech

The CNMF reverberation model was modified by incorporating a model for clean speech. The clean speech spectrogram was approximated using a NMF model as,

$$\mathbf{S} \approx \mathbf{W}_s \mathbf{X}_s$$
$$S(k,n) \approx \sum_{r=1}^{R_s} W_s(k,r) X_s(r,n), \tag{4.4}$$

where $\mathbf{W}_s \in \mathbb{R}_+^{K \times R_s}$ and $\mathbf{X}_s \in \mathbb{R}_+^{R_s \times (T-L_h+1)}$ represents the bases and activation matrix. R_s represents the rank of NMF decomposition. $W_s(k,r)$ and $X_s(r,n)$ are elements of $\mathbf{W}_s$ and $\mathbf{X}_s$, respectively. Utilizing the NMF model in (4.1), the reverberation model can be rewritten as,

$$Y_R(k,n) \approx \sum_{l=0}^{L_h} H(k,l) \Big[\sum_{r=1}^{R_s} W_s(k,r) X_s(r,n-l) \Big] \tag{4.5}$$

An algorithm was proposed to obtain clean speech and RIR spectrograms from reverberation model in (4.5) [25, 26]. The optimization problem can be summarized as,

$$C_{cnmf1} = \underset{\mathbf{H},\, \mathbf{W}_s,\, \mathbf{X}_s}{\operatorname{argmin}} \Big[\sum_{k,n} \mathrm{KL}\Big(Y_R(k,n) | \tilde{Y}_R(k,n) \Big) + \lambda_{cnmf1} \sum_{r,n} X_s(r,n) \Big]$$

Subjected to

$$H(k,n) \geq 0,\ W_s(k,r) \geq 0,\ X_s(r,n) \geq 0$$
$$H(k,0) = 1 \forall k \in \{0, 1, ..., (K-1)\}, \tag{4.6}$$

where $\tilde{Y}_R(k,n)$ represents the estimated reverberated spectrogram based on reverberation model in (4.5). λ_{cnmf1} represents a weighting factor. The objective function has two terms. The first term minimizes the modeling error between estimated and actual reverberated speech spectrograms. The second term introduces sparsity in estimated clean speech activation. The first set of constraints ensures that the estimated clean speech and RIR spectrograms are non-negative. Normalization in the RIR spectrogram removes the inherent scaling ambiguity present in the cost function.

Multiplicative update rule was obtained for solving the optimization problem in (4.6) [25, 26]. The update rules are summarized in (4.7).

$$
\begin{aligned}
H(k,n) &\leftarrow H(k,n)\frac{\sum\limits_{l}\dfrac{Y_R(k,l)}{\tilde{Y}_R(k,l)}S(k,n-l)}{\sum\limits_{l}S(k,n-l)}\\
W_s(k,r) &\leftarrow W_s(k,r)\frac{\sum\limits_{n,l}\dfrac{Y_R(k,n)}{\tilde{Y}_R(k,n)}H(k,l)X_s(r,n-l)}{\sum\limits_{n,l}H(k,l)X_s(r,n-l)}\\
X_s(r,n) &\leftarrow X_s(k,r)\frac{\sum\limits_{k,l}\dfrac{Y_R(k,l)}{\tilde{Y}_R(k,l)}H(k,n-l)W_s(k,r)}{\sum\limits_{k,l}H(k,n-l)W_s(k,r)+\lambda_{cnmf1}}
\end{aligned}
\tag{4.7}
$$

where $\tilde{S}(k,n) = \sum\limits_{r} W_s(k,r)X_s(r,n)$. There exists two approaches. This distinction is based on how $W_s(k,r)$ is estimated. In unsupervised (offline) approach, the clean speech bases $W_s(k,r)$ is estimated from the reverberated data. This approach is referred to as CNMF1_u. In supervised (online) approach, $W_s(k,r)$ are pre-learned. A NMF decomposition is performed on the spectrogram of available clean speech recording. The bases vectors obtained for this decomposition forms $W_s(k,r)$. This method is referred to as CNMF_s. The steps involved are summarized in Algorithm 2.

Algorithm 2: Steps involved in CNMF1_s and CNMF1_u

Result: Enhanced speech spectrogram $S(k,n)$

initialize $W_s(k,r)$, $X_s(r,n)$, $H(k,n)$ to random positive values;

for $i = 1 : i\_max$ **do**

  if $W_s(k,r)$ *is not fixed* **then**

    update $W_s(k,r)$

  end

  update $X_s(r,n)$

  update $H(k,n)$

  if $H(k,n) > H(k,n-1)$ **then**

    truncation of $H(k,n)$

    $H(k,n) \leftarrow \min\Big(H(k,n), H(k,n-1)\Big)$

  end

  normalization $H(k,n) \leftarrow \dfrac{H(k,n)}{H(k,0)}$

end

estimation of clean speech spectrogram $S(k,n) = \dfrac{\sum_{r=1}^{R_s} W_s(k,r)X_s(r,n)}{\tilde{Y}_R(k,n)} Y_R(k,n)$

4.1.3 Comparison of reverberation models

4.2 Incorporation of RIR properties on optimization problem

The NMF based dereverberation methods discussed in Section 4.1 used properties of clean speech spectrogram like low-rank nature and sparsity. However, such methods did not incorporate any properties of RIR except for basic truncation and normalization of the RIR spectrogram. This chapter focuses on incorporating three meaningful constraints on the RIR spectrogram to improve the dereverberation performance of basic NMF based dereverberation methods. The properties of RIR used are (i) frequency envelope of RIR spectrogram, (ii) sparsity of RIR spectrogram, and (iii) retaining the early part of the RIR spectrogram. These approaches are explained in detail next.

4.2.1 Frequency envelope of RIR spectrogram

4.2.2 Sparsity of RIR spectrogram

4.2.3 Retaining early part of RIR

Chapter 5

Separability Assumption on RIR Spectrogram

5.1 NMF degradation models

- Derivation
  - NMF model for reverberation
  - Extended model for reverberation and noise

5.2 Analysis of model

- Effect of reverberation on clean speech bases and activation

5.3 Enhancement algorithm

- cost function
- multiplicative update rule
- normalization used

5.4 Experimental results

- enhancement results
- RIR estimates

5.5 Discussion

- comparison of results
- limitations

Chapter 6

Low-rank NMF model for RIR spectrogram

6.1 Justification of low-rank approximation

- Quality of approximation
- generalization of separability approximation
  - remove limitations of earlier work
  - easy to interpret model

6.2 NMF model for degraded spectrogram

- derivation of NMF model for reverb and degraded spectrogram
- comparison of degradation model obtained with proposed method when compared with NMF based methods in literature
- effect of clean speech bases and activation with reverberation

6.3 Algorithm details

- cost function
- multiplicative update rule

Method	WER (%)
Clean	3.44
Degraded	62.34
CNMF0	47.95
CNMF1_s	44.45
CNMF2_s	43.24
RNMF_s	**40.24**

Table 6.1: Enhancement results when reverberated with RIR RIR3_far and stationary noise added with 10 dB SNR. The RIR has $T_{60} \approx 700$ ms and source-microphone distance of 2 m.

Method	**WER (%)**			
Clean	3.44			
Degradation condition	**R2_near**	**R2_far**	**R3_near**	**R3_far**
		48.57	19.41	62.34
CNMF0				47.95
CNMF1_s				44.45
CNMF2_s				43.24
RNMF_s				**40.24**

Table 6.2: Enhancement results for 10 dB SNR stationary noise.

6.4 Experimental results

- enhancement results, ASR results
- variation of performance with RIR, SNR, rank of decomposition, etc.

6.5 Discussion

Method	WER (%)			
Clean	3.44			
Degradation condition	**R2_near**	**R2_far**	**R3_near**	**R3_far**
CNMF0				
CNMF1_s				
CNMF2_s				
RNMF_s				

Table 6.3: Enhancement results for 20 dB SNR stationary noise.

Method	WER (%)			
Clean	3.44			
Degradation condition	**R2_near**	**R2_far**	**R3_near**	**R3_far**
CNMF0				
CNMF1_s				
CNMF2_s				
RNMF_s				

Table 6.4: Enhancement results for noise free condition.

Appendix A

Supporting Material

References

[1] K. Kinoshita *et al.*, “A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[2] N. Patrick and G. Nikolay, *Speech Dereverberation*. New York: Springer, 2010.

[3] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth’chime’speech separation and recognition challenge: Dataset, task and baselines,” *arXiv preprint arXiv:1803.10609*, 2018.

[4] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third âĂŸCHiMEâĂŹspeech separation and recognition challenge: Dataset, task and baselines,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[5] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” vol. 15, no. 7, pp. 2011–2022, 2007.

[6] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[7] T. Nishiura, S. Nakanura, and K. Shikano, “Speech enhancement by multiple beamforming with reflection signal equalization,” in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 1. IEEE, 2001, pp. 189–192.

[8] G. Xu, H. Liu, L. Tong, and T. Kailath, “A least-squares approach to blind channel identification,” *IEEE Transactions on signal processing*, vol. 43, no. 12, pp. 2982–2993, 1995.

[9] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on signal processing*, vol. 51, no. 1, pp. 11–24, 2003.

[10] M. I. Gurelli and C. L. Nikias, "Evam: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals," *IEEE Transactions on Signal Processing*, vol. 43, no. 1, pp. 134–149, 1995.

[11] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, p. 769285, 2003.

[12] Y. A. Huang and J. Benesty, "Adaptive multi-channel least mean square and newton algorithms for blind channel identification," *Signal Processing*, vol. 82, no. 8, pp. 1127–1138, 2002.

[13] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, 1979.

[14] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7. IEEE, 1982, pp. 1858–1861.

[15] B. D. Radlovic and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 728–737, 2000.

[16] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 2, pp. 145–152, 1988.

[17] H. Yamada, H. Wang, and F. Itakura, "Recovering of broadband reverberant speech signal by sub-band mint method," in *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1991, pp. 969–972.

[18] P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Inverse filter design and equalization zones in multichannel sound reproduction," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 3, pp. 185–192, 1995.

[19] K. Lebart, J.-M. Boucher, and P. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[20] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech dereverberation based on variance-normalized delayed linear prediction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[21] H. Buchner, R. Aichner, and W. Kellermann, “TRINICON-based blind system identification with application to multiple-source localization and separation,” *Blind speech separation*, pp. 101–147, 2007.

[22] H. Kameoka, T. Nakatani, and T. Yoshioka, “Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 45–48.

[23] Y. Avargel and I. Cohen, “System identification in the short-time Fourier transform domain with crossband filtering,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[24] K. Kumar, R. Singh, B. Raj, and R. Stern, “Gammatone sub-band magnitude-domain dereverberation for ASR,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[25] N. Mohammadiha and S. Doclo, “Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 2, pp. 276–289, 2016.

[26] N. Mohammadiha, P. Smaragdis, and S. Doclo, “Joint acoustic and spectral modeling for speech dereverberation using non-negative representations,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[27] S. Wager and M. Kim, “Collaborative speech dereverberation: Regularized tensor factorization for crowdsourced multi-channel recordings,” in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1532–1536.

[28] S. Mirsamadi and J. H. L. Hansen, "Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for ASR applications," in *Proc. Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[29] H. Kallasjoki, J. F. Gemmeke, K. J. Palomaki, A. V. Beeston, and G. J. Brown, "Recognition of reverberant speech by missing data imputation and NMF feature enhancement," in *Proc. REVERB Workshop*, May 2014.

[30] X. Li, S. Gannot, L. Girin, and R. Horaud, "Multichannel identification and nonnegative equalization for dereverberation and noise reduction based on convolutive transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1755–1768, 2018.

[31] D. Baby and H. V. Hamme, "Supervised speech dereverberation in noisy environments using exemplar-based sparse representations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 156–160.

[32] D. Baby, "Non-negative sparse representations for speech enhancement and recognition," Ph.D. dissertation, University of Leuven, 2016.

[33] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[34] J. Y. Wen, E. A. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 329–332.

[35] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant, and P. Vary, "Do we need dereverberation for hand-held telephony?" in *Proc. Int. Congress on Acoustics (ICA), Sydney, Australia*, 2010.

[36] N. Mohanan, R. Velmurugan, and P. Rao, "A non-convolutive NMF model for speech dereverberation," in *Proc. INTERSPEECH*, 2018, pp. 1324–1328.

[37] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[38] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. OâĂŹBrien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

List of Publications

Put your publications from the thesis here. The packages `multibib` or `bibtopic` or `biblatex` or enumerate environment or thebibliography environment etc. can be used to handle multiple different bibliographies in the document.

Acknowledgements

This section is for the acknowledgments. Please keep this brief and resist the temptation of writing flowery prose! Do include all those who helped you, e.g. other faculty/staff you consulted, colleagues who assisted etc.

Nikhil M
IIT Bombay
5 May 2020