ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

SPARSITY BASED FORMULATIONS FOR DEREVERBERATION

M.Sc. THESIS

Aziz KOÇANAOĞULLARI

Electronics and Communication Engineering Department

Telecommunications Engineering Programme

MAY 2016

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE
ENGINEERING AND TECHNOLOGY

SPARSITY BASED FORMULATIONS FOR DEREVERBERATION

M.Sc. THESIS

Aziz KOÇANAOĞULLARI
(504141303)

Electronics and Communication Engineering Department

Telecommunications Engineering Programme

Thesis Advisor: Assoc. Prof. Dr. İlker BAYRAM

MAY 2016

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

YANKILAŞIM GİDERMEK İÇİN SEYREKLİK TABANLI DÜZENLEMELER

YÜKSEK LİSANS TEZİ

Aziz KOÇANAOĞULLARI
(504141303)

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Telekomünikasyon Mühendisliği Programı

Tez Danışmanı: Assoc. Prof. Dr. İlker BAYRAM

MAY 2016

Aziz KOÇANAOĞULLARI, a M.Sc. student of ITU Graduate School of Science Engineering and Technology 504141303 successfully defended the thesis entitled **"SPARSITY BASED FORMULATIONS FOR DEREVERBERATION"**, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Assoc. Prof. Dr. İlker BAYRAM** ....
Istanbul Technical University

Jury Members : **Prof. Dr. Bilge GÜNSEL KALYONCU** ....
Istanbul Technical University

Assoc. Prof. Dr. A. Taylan CEMGİL ....
Boğaziçi University

Date of Submission :2 May 2016
Date of Defense : 7 Jun 2016

To my beloved parents and brother,

FOREWORD

Special thanks to my supervisor İlker BAYRAM who has never spared his help, attention and valuable time, for providing me such an opportunity for thesis.

During my thesis I received financial support from TUBITAK as a scholarship student with project number 113E511.

Special thanks to my entire family, especially to my mother, father and brother for the unending support and encouragement.

May 2016 Aziz KOÇANAOĞULLARI

TABLE OF CONTENTS

Page

ABBREVIATIONS

DR	**:** Douglas-Rachford (Algorithm)
FT	**:** Fourier Transform
IFT	**:** Inverse Fourier Transform
ISTA	**:** Iterative Shrinkage Thresholding Algorithm
LSE	**:** Least Squares Estimate
RIR	**:** Room Impulse Response
STFT	**:** Short Time Fourier Transform

LIST OF TABLES

LIST OF FIGURES

Page

SPARSITY BASED FORMULATIONS FOR DEREVERBERATION

SUMMARY

Acoustic signals recorded in concerts, meetings or conferences are effected by the room impulse response and noise. Estimating the clean source signals from the observations is referred as the dereverberation problem. If the room impulse responses are known, the problem is non-blind dereverberation problem. In this thesis non-blind dereverberation problem is posed using convex penalty functions, with a convex minimization procedure. The convex minimization problems are solved using iterative methods. Through the thesis sparse nature of the time frequency spectrum is referred. In order to transform the time domain signal to a time frequency spectrum Short Time Fourier Transform is used.

In the thesis, to begin with, the general problem is defined in time domain. The basics of the dereverberation is proposed. Basics of the convex minimization procedure is explained. Douglas Rachford Algorithm which is used to solve complicated convex minimization problems is explained.

The chapter 3 proposes a derevereberation formulation based on sparsity. The dereverberation problem, with known room impulse response, is conventionally posed as a sparsity based minimization problem, by masking Short Time Fourier Transform coefficients. The sparsity constraint can be posed using an ℓ_1 norm type penalty function. However, in such formulations, especially if the room impulse response is longer than the windowing function of the Short Time Fourier Transform the reverberation effects can not be directly represented in the transform domain. Therefore, the minimization iterations require transform and its inverse in order to mask Short Time Fourier Transform coefficients after time domain deconvolution. Changing domains is more time consuming compared to masking STFT coefficients, in turn increases computational time dramatically. In order to get rid of the transformation requirement, the room impulse response is represented in Short Time Fourier Transform frequency bands. With the approximation filters room impulse response is denoted as a convolutive operator in each frequency band. In this chapter an algorithm proposed, that does not require Short Time Fourier Transform and its inverse, using the proposed approximates of the room impulse response. Also the room impulse response approximation and the dereverberation with the sparsity constraint are justified with experiments. Experiments show that, sparsity based solution yields musical noise.

In the chapter 4, musical noise is suppressed using phase information of the coefficients in a frequency band. It can be observed that in a frequency band, time consecutive coefficients are active through a harmonic. These coefficients tend to have close magnitudes. In addition, phase shift between coefficients in harmonics can be considered as constant and phase information is unimportant outside the harmonics.

It can be considered that for each harmonic, there lies a complex number that maps time consecutive coefficients together. Outside the harmonics the matching constant is tend to be 0. Therefore, a piece wise constant mask can be found that maps a frequency band to its own phase shifted version. This mask, in fact, satisfies the sparsity property, as it is mapping a sparse frequency band to another. In this chapter, a method for estimating the mask is proposed. The mask is applied on least squares estimate of the signal. The least squares estimate can be performed with known impulse response and noise properties. The dereverberation performance is justified using experiments. Through the experiments different audio signals with different input signal to ratio values are taken into consideration. Also, different weight compositions are taken into consideration.

In the chapter 5, different from the chapters 3 and 4, multiple microphone case is taken into consideration. Multichannel case is often solved making use of microphone array geometry or using multi channel penalty functions. Another important struggle in dereverberation is estimating RIRs. Instead of measuring explicit RIRs for the observations, a common filter can be obtained using preliminary observations. This information can be exploited in multichannel estimation. In this chapter a minimization procedure with a multi channel penalty function is proposed. In multiple microphone case, the observations share a common information. A time frequency coefficient is expected to be active in all the observations, if the microphones are close. In order to make use of that information the estimation can be modified into multi channel estimation. Instead of estimating the source signal from observations, it is assumed that the observations are formed from different sources. The sources are defined as shortly reverberated versions of the source signal. In order to obtain these observations, relatively short room impulse response definition is required. With the definition the room impulse response can be divided into two: the common part, which is the same for all observations and the independent part, which differs with position. Thus, a formulation for mixed norm is proposed using relatively short impulse responses. However, this algorithm can be generalized. In order to relax the condition on time frequency coefficients, it can be assumed that time shift between harmonics between observations is relatively small. Thus the harmonic structure is investigated using blocked mixed norm regularization. Both algorithms for mixed norm and blocked mixed norm regularization are justified and compared using experiments on speech signals.

YANKILAŞIM GİDERMEK İÇİN
SEYREKLİK TABANLI DÜZENLEMELER

ÖZET

Konser, konferans, toplantı gibi ortamlarda kaydedilen akustik işaretler, kaydın alındığı ortam nedeni ile yankıya ve gürültüye maruz kalır. Kaynak işaretinin elde edilen gözlemlerden kestirimi yankı giderme problemi olarak isimlendirilir. Bu kayıtlarda göze çarpan yankı etkileri bir süzgeç olarak zaman tanım bölgesinde modellenebilir. Yankı etkilerini modelleyien bu süzgeç oda darbe cevabı olarak isimlendirilir. Oda darbe cevabının bilindiği durumda problem gözü kapalı olmayan yankı giderme problemine dönüşür. Tez boyunca oda darbe cevabının bilindiği durumlar dikkate alınmıştır. Gözlemlenebilir ki, oda darbe cevabı kaynak ve gözlem noktalarına çok bağımlıdır. Bu nedenle oda darbe cevabının bütün uzaydaki noktalar için kestirimi çok zordur. Bu durumda oda darbe cevapları tezdeki deneylerde sentetik olarak uygulanmış veya gözlem ortamında kayıt alındığı sırada gözlemden elde edilmişlerdir. Bölüm 5, bu duruma farklı bir açıdan bakılmasının örneğidir. Bu bölümde oda darbe cevabının kısmen bilindiği ve gözlem ortamı için tek bir süzgeç tanımlanabileceği durumları göz önüne alınmıştır.

Bu tezde gözü kapalı olmayan yankı giderme problemi, dışbükey bir en küçükleme problemi yardımıyla çözülmüştür. Dışbükey en küçükleme problemleri yinelemeli yöntemler kullanılarak çözülmüştür. Tez boyunca, farklı ceza terimleri kullanılmış olsa da, ceza terimleri, işaretin zaman sıklık dönüşümü altında seyrek yapıya sahip olacağını varsaymaktadır. Seyreklik koşulundan anlaşılması gereken, sayılı zaman sıklık katsayısının aktif olduğu ve aktif katsayı kümelerinin uzaya dağılmış olduğudur. Seyrek zaman sıklık dönüşümü katsayıları yapısı kullanılarak kestirim daha iyi bir biçimde sağlanabilir. Zaman sıklık dönüşümü olarak Kısa Zamanlı Fourier Dönüşümü kullanılmıştır. Önerilen dönüşüm yerine herhangi bir doğrusal zaman sıklık dönüşümü de kullanılabilir.

Tezde ilk olarak, zaman tanım bölgesinde, genel gözlem modeli verilmiştir. Bu yolla, yankı giderme probleminin temelleri anlatılmıştır. Dışbükey en küçükleme modeli verilmiştir. Karmaşık en küçükleme problemlerini çözmek amacıyla yinelemeli bir yöntem olan Douglas Rachford Algoritması açıklanmıştır. Bu algoritma gradyan hesaplaması kolay olmayan problemleri, hesabı kolay olan iki alt probleme ayırarak yinelemeli olarak çözmektedir. Yinelemelerde her bir verilen alt problemlerin kısıtlamaları, bu fonksiyonların yakınsal terimleri kullanılarak sağlanmaktadır. Bu yinlemelerde, belirlenmiş olan bir adım değeri ile kısıtlara yaklaşılır. En son olarak ise bu kısıt kümeleri arasındaki en yakın nokta bulunarak en uygun noktaya ulaşılmış olunur.

Bölüm 3 seyreklik koşulu altında yankı giderme problemine ayrılmıştır. Geleneksel yöntemler, yankı giderme problemini, oda darbe cevabı bilindiği durumda, Kısa Zamanlı Fourier Dönüşümü katsayılarını maskeleyerek çözmektedir. Seyreklik koşulu

dönüşüm katsayıları üzerine ℓ_1 normlu bir ceza terimi kullanılarak sağlanmaya çalışılmıştır. Ancak, bu tip önermelerde oda darbe cevabının, Kısa Zamanlı Fourier Dönüşümü alınırken kullanılan pencereden uzun olması durumunda, oda darbe cevabu, dönüşüm tanım bölgesinde çarpım olarak ifade edilemez. Seyreklik koşulunun dönüşüm katsayıları üzerinde arandığı ve oda darbe cevabının zaman tanım bölgesinde bir evrişim işleci olarak tanımlandığı düşünülürse, en küçükleme probleminin yinelemeleri sırasında tanım bölgesini değiştirmeye ihtiyaç olacaktır. Tanım bölgesi değiştirmenin, maskeleme işlemine göre daha uzun süreceği açıktır. Bu nedenle tanım bölgesi değiştirmek en küçükleme probleminin çözüm sürecini çok arttırmaktadır. Bu etkiden kurtulmak amacıyla oda darbe cevabının etkileri dönüşüm katsayılarının frekans bantları için bir evrişim işleci olarak ifade edilmiştir. Bu bölümde yankı giderme işlemi için yinelemelerinde yanım bölgesi dönüşümü olmayan bir algoritma önerilmiştir. Aynı zamanda deneyler ile dönüşüm tanım bölgesinde yaklaşık olarak elde edilen oda darbe cevabı süzgeçlerinin ve yankı giderme algoritmasının başarımı tartışılmıştır. Deneyler sonunda, seyreklik koşulu ile yankı giderme işlemi sonucunda müzik gürültüsünün ortaya çıktığı görülmüştür.

Bölüm 4, faz bilgisi kullanılarak, müzik gürültüsünün azaltılmasına ayrılmıştır. Seçili herhangi bir sıklık bandında, harmonikler üzerinde zamanda ardışık gelen katsayıların etkin olduğu gözlemlenebilir. Bu katsayıların genlik değerleri birbirine yakındır, aynı zamanda iki katsayı arasında faz kaymasının yaklaşık sabit olduğu söylenebilir. Bu yolla her harmonik için faz kaymasını modelleyecek bir karmaşık sayı bulunabilir. Harmonikler dışında, işaret zaman sıklık katsayılarının 0a yakın genlikli olması beklendiği için, faz bilgisi anlamsızlaşır. Bu yolla parça başı sabit bir maske ile bir sıklık bandı, kendisinin fazı kaymış biçimine bağlanabilir. Bu maske, aynı zamanda dönüşüm katsayılarının özelliklerini de korumaktadır. Bir dönüşüm sıklık bandını bir diğerine bağlaması nedeniyle seyrek bir yapıda olması da beklenir. Bu bölümde, her sıklık bandı için tanımlanan maskelerin kestirimi için bir en küçükleme problemi önerilmiştir. Bulunan maskeler, en küçük kareler kestirimi sıklık bantlarına uygulanmıştır. En küçük kareler kestirimi, oda darbe cevabı ve gürültü özellikleri bilindiği varsayımında, rahatlıkla hesaplanabilir. Deneylerde farklı ses işaretleri kullanılarak önerilen yöntemin başarımı tartışılmıştır. Aynı zamanda kestirim başarımları farklı seyreklik, faz sabitliği ağırlıkları için sunulmuştur.

Bölüm 5, 3 ve 4 bölümlerinden farklı olarak çok gözlemin olduğu duruma ayrılmıştır. Çok mikrofon ile gözlem elde edilen durumlarda, yankı giderme problemi genel olarak mikrofon yerleşimleri kullanılarak veya çok kanallı ceza terimleri yardımıyla çözülür. Yankı giderme probleminde temel zorluklardan biri de oda darbe cevaplarının bulunmasıdır. Bu süzgeçlerin bulunması yerine, kayıt ortamı için ortak bir süzgeç tanımı yapılabilir ve bu tanımla çok kanallı en küçükleme ortaya atılabilir. Bu temel süzgecin belirlenmesi gözlem anından önce yapılan deneylerle mümkün olabilir. Temel süzgeçte öne atılan sav, oda darbe cevaplarının ortak bir noktalarının bulunduğudur. Bu ortak nokta ise tüm gözlem ortamı için geçerlidir. Bu tanım kullanılarak elde edilen temel süzgeç ve ardından elde edilen süzgeç artıkları yardımı ile yankı etkilerinin kısaltılması sağlanabilir. Özellikle temel süzgecin oda özelliklerini yansıttığı varsayıldığında, bu süzgecin gözlemlerden silinmesi yankıyı önemli ölçüde ortadan kaldıracaktır. Bu bölümde, yankı giderme problemini çözmek amacıyla, çok kanallı bir ceza terimi ortaya atılmıştır. Ceza terimi kullanılarak bir dışbükey en küçükleme problemi ortaya konmuştur. Çok mikrofonlu kayıtlarda gözlemlerde ortak bilgi göze çarpar. Bir zaman sıklık katsayısının, yaklaşık olarak tüm gözlemlerde etkin olması beklenir. Bu durum zaman sıklık katsayıları faklı gözlemler için üst üste konumlandırıldıklarında görülebilir. Bu bilgiden faydalanmak amacıyla, dışbükey en küçükleme problemi yardımıyla kaynak işareti kestirimi yerine, çok kanallı kestirim yoluna gidilmiştir. Bu nedenle gözlemlerin farklı kaynaklardan elde edildiği varsayılmıştır. Farklı kaynaklar, kaynak işaretinin kısa biçimde yankılanmış biçimleri olarak düşünülmüştür. Bu kaynakları tanımlamak amacıyla görsel olarak kısalmış oda darbe cevabı tanımı yapılmıştır. Bu tanımdan faydalanılarak karma norm ceza terimi kullanılarak bir en küçükleme problemi ortaya atılmıştır. Önerilen yöntem ceza teriminin rahatlatılması ile genelleştirilebilir. Zaman sıklık katsayılarındaki mikrofon üzerinden seyreklik araştırması yerine, harmonik araştırması yapılabilir. Bu amaçla bu bölümde kümelenmiş karma norm ceza terimi de ortaya atılmıştır. Bu terimin ortaya atılmasındaki temel itici güç ise zaman sıklık katsayılarının gürültü etkisinde her gözlemde bulunmayacağının garantisinin verilemeyeceğidir. Katsayıların kümelenmesi yardımıyla bu gürültü terimlerinin ardışıl gelme olasılığı düşürülerek daha temiz bir kestirim elde edilebilir. Her iki yöntem de deneyler bölümünde sınanmış ve karşılaştırılmıştır.

Son bölümde ise tezde anlatılan konular tekrardan göz önüne serilmiştir. Bu serimlerden yola çıkarak, tezden yararlanarak hangi konularda araştırma yapılabileceğine de yer verilmiştir.

1. INTRODUCTION

Audio signals recorded in conferences, meetings or concerts are effected by the properties of the environment and the noise. Especially in recordings with microphones, distant from the source or within a class or hall, the observations are reverberated. This reverberation effect can be modeled with a linear operator. The impulse response of that operator model is named as room impulse response. With the presence of noise or other sources, the deteriorated signal is hard to understand for human auditory system. In many possible applications like seminar or public meeting recordings the aim is to obtain the speaker as clean as possible. In order to revert the effects of reverberation, methods address the deconvolution problem. The problem to estimate source signal from reverberated and noisy observation is referred as the dereverberation. The problem model and some methods can be observed from [1].

In this thesis dereverberation problem with known impulse response is taken into consideration. Therefore, it is called non blind dereverberation. The dereverbration problem is posed as a convex minimization problem with variety of penalty functions. In chapter 3 and chapter 4 single channel dereverberation is performed. In chapter 5 multiple observations are taken into consideration.

In the chapter 2, the system model is proposed. The convex optimization procedure is explained. Douglas Rachford algorithm, which is an iterative solution for convex optimization problems, is explained. DR algorithm requires proximity operators of functions. The proximity of a function is defined and for some fundamental functions, proximals are calculated. After explaining the basics used through this thesis, the methods for the dereverberation is proposed.

It is conventional, in acoustic source estimation, to assume that time frequency transformation of a signal is sparse. Thus, derevereberation can be performed with a sparsity constraint on the estimate spectrum. In the chapter 3, a sparsity based deconvolution problem in STFT domain is proposed. Least squares term is used in order to minimize the error between the estimate and the original signal, linking them with the knowledge of RIR. This process is conventionally done in time domain because RIR is a convolutive operator defined in the time domain. However, the penalty function is required to enforce the sparsity of the STFT coefficients. As a result in the minimization procedure, it is required to use transform domain for sparsity constraint and the time domain for deconvolution. It can be observed that, pursuing penalty functions in different domains requires domain changes in calculations. This point of view becomes more problematic if the solution is obtained using an iterative method, as domain change is required at each iteration. In the steps, it is required to use transform and its inverse at least once for each iteration. In order to avoid domain changes, RIR is required to be represented in the transform domain. Therefore, a room impulse response model in STFT domain is proposed. Such that impulse response model provides the freedom of using convex constraints in STFT frequency bands without increasing computational complexity. Dereverberation with sparsity constraint is a common method, however, with proposing a method in transform domain, the computational time is dramatically decreased. Using proposed method, dereverberation is achieved with similar results to conventional methods. The negative result of sparsity based dereverberation is the musical noise caused by high frequency coefficients. These effects can be canceled by increasing the weight of sparsity constraint. However, some harmonics are lost after that modification, which results in decreased quality in estimation. In order to get rid of the musical noise without losing any information, a new method is required.

In the chapter 4, a dereverberation formulation employing phase information is proposed as a continuation of the first chapter. It is notified that, sparsity based denoising methods reduce the noise by modifing magnitudes. However with the effects of the reverberation the sparse nature of the coefficients is questionable and these methods also yield musical noise. The musical noise caused by conventional sparsity based methods can be erased using phase information of the time-frequency coefficients of the transform. The phase information is used to relate the coefficients in a harmonic. In a harmonic of a specified frequency band, consecutive time coefficients tend to have similar magnitude value and a constant shift between coefficients. Phase information of the coefficients are highly affected by noise and reverberation, however the phase shift manages to stay close to constant. In harmonics, with the usage of constant phase shift property, consecutive coefficients can be related. Under the assumption that phase shift constancy holds, a complex number can be found for each harmonic that maps consecutive coefficients. Also this can be generalized such that, the complex number also maps the harmonic to phase shifted version of itself. Assuming there exists few harmonics in a frequency band, a piecewise constant mask can be formed. This mask maps the signal to phase shifted version of itself. This mask preserves harmonic structure. An algorithm employing this mask is proposed to erase musical noise while preserving the harmonics. As stated before the algorithm employs phase information of the coefficients in addition to magnitude information. The quality of the proposed algorithm is justified and the improvement achieved is compared to sparsity based estimate through experiments.

The methods proposed in chapters 3 and 4, are using a single microphone. In applications microphone arrays are used in order to increase the efficiency where there is a set of observations. Different from the previous chapters in the chapter 5 multichannel dereverberation is taken into consideration. Multi channel dereverberation problem is conventionally solved expoiting the microphone array geometry or multi channel penalty functions. Therefore, using a multichannel convex penalty function, this property can be exploited in order to obtain the estimate. Using different RIR for each observation point with one source is challenging. In order to prevent using different room impulse responses the shortened impulse response concept is explained. It is assumed that room impulse responses of microphones share a common and shorter part. Thus in the problem it is assumed that the common part is the filter and the observations are obtained using different sources with the same information. Therefore, the time frequency coefficients of these sources can be linked together. Assuming that sources are located close enough, corresponding to the geometry of the microphone array, it is acceptable to assume their time frequency spectrogram can be related. With the assumption that, microphones are closely located, it is expected that shift in time domain observations is considerably little compared to the windowing function employed in STFT. Therefore in a frequency band the harmonics are expected to be closely located. This property can be exploited using multiple observations. In order to use that information, with the usage of shortened impulse response definition, shortly reverberated observations are considered as the sources. Instead of estimating the source, the problem is modified to address estimating shortly reverberated observations. In these observations, it is expected that if a time frequency coefficient is active in one spectrum, it is also active in others. With that information, while investigating the sparsity of the transform spectrum, all observation spectra is taken into consideration. With a mixed norm penalty function the sparsity of a a frequency band is enforced through all observation bands. An algorithm with mixed norm penalty function defined. However, harmonics in shortly reverberated observations are closely located. Also number of harmonics is limited. This leads to grouping time frequency coefficients in a frequency band and investigating presence of harmonic using all observation spectra. If the constraint proposed with mixed norm is relaxed, a block mixed norm algorithm exploiting that information is proposed for dereverberation. Also with assuming RIRs to be δ functions, problem becomes the denoising problem. The denoising and dereverberation performances of the method are justified separately using experiments.

In the final chapter conclusions and remarks are given. Also possible future research is proposed.

All these methods are sparsity based dereverberation methods which can be considered a common point. Sparse nature of time frequency spectrum of the audio signals motivates defining sparsity based convex optimization problems. Also, with the convergence of various algorithms applicable on such problems, different constraints can be applied. With the motivation of proposing different methods for sparsity based dereverberation, the performance of the estimates are tried to be improved. The aim of this thesis is to pose various convex methods for both single channel and multichannel data that can improve the efficiency and quality of source signal estimation.

2. PRELIMINARY CHAPTER

In this chapter the dereverberation concept is tried to be explained. Additionally Douglas Rachform Algorithm and proximal definitions are made. These definitions are also remarked in other chapters if required.

The audio recording setup is demonstrated in Fig.2.1. The arrows between microphones and speakers denote the reverberation effects, this setup can be modeled,

$$y_a = \sum_{b=1}^{B} h_{ab} * x_b + n_a \quad a = 1, \cdots, A \tag{2.1}$$

where $*$ denotes the convolution operator. There y's denote the observations, x's denote the source signals, h's denotes the room impulse responses that cause reverberation and n's denote the channel noise. As demonstrated in the setup for a microphone all the sources has their own filters. It can be observed that RIR changes dramatically with position. Here in Eq.(2.1), with that notation, the difference of RIR with respect to location is modeled. This time domain model is used through the thesis.

In the thesis the convolution with RIR is denoted with the H in order to preserve simplicity. Thus the problem defined in Eq.(2.1) becomes,

$$y_a = \sum_{b=1}^{B} H_{ab} x_b + n_a \quad a = 1, \cdots, A \tag{2.2}$$

The room impulse responses are usually of few [s] length. Thus, with the reverberation, each activity in time domain is expected to be extended. The reverberation effects are visualized in Fig.2.2. As can be observed from the Fig.2.2 the observation is extended. With the noise added, auditory quality decreases. The aim is to estimate original signal from the reverberated observations. For example, from the noisy and reverberated observation in Fig.2.2-(c), the aim is to estimate the original signal in Fig.2.2-(a).

2.1 System Model

In this thesis the RIR relations between microphones and sources are assumed to be known. Thus this poses the non blind deconvolution problem, where the only

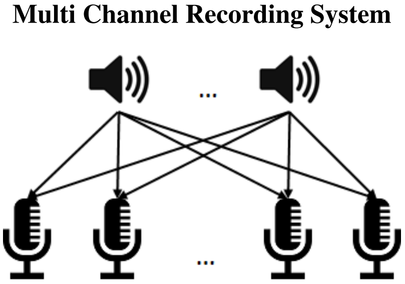

Figure 2.1: The general setup used for multi channel audio recording with multiple sources.

unknown is the source signal. On the other hand, if the channel filters are required to be estimated, the problem becomes blind deconvolution problem. However, this problem is ill posed as it requires two concepts to be estimated at the same time [2–4]. Even with the known RIR, the estimate can be improved using some properties. The most common property exploited is sparsity of the spectrum.

In order to employ this property time frequency transformation is required. In this thesis short time Fourier transform is used. Assuming that $g(n)$ is a low pass filter and $g_{k,l}(n) = g(n - l\Delta_t)\exp(-jk\Delta_w(n - l\Delta_t))$, where Δ_w denote the frequency length and Δ_l denotes the shift size of the window function. The STFT of a signal is calculated with,

$$X(k,l) = \langle x(n), g_{k,l}(n) \rangle \tag{2.3}$$

Where $\langle \cdot, \cdot \rangle$ denotes the inner product. It can be observed that if a filter has a length less than or equal to the windowing function, the effects caused by that filter can be denoted as an elementwise multiplication in STFT domain. This, in case, can be explained that the corresponding coefficients of the signal and room impulse response lie in the same interval. Thus the convolution can be represented as elementwise multiplication in these intervals. This can be proven in short, using the definition of Fourier transform. Convolution in time domain corresponds to multiplication in frequency domain. However, as can be seen from Fig.2.2-(b) the room impulse response length is considerably long. In order to obtain acceptable resolution, the window length in STFT for audio applications is chosen around 50[ms], where the room impulse response is of length 1[s]. Thus the reverberation can not be represented as an elementwise multiplication in STFT domain.

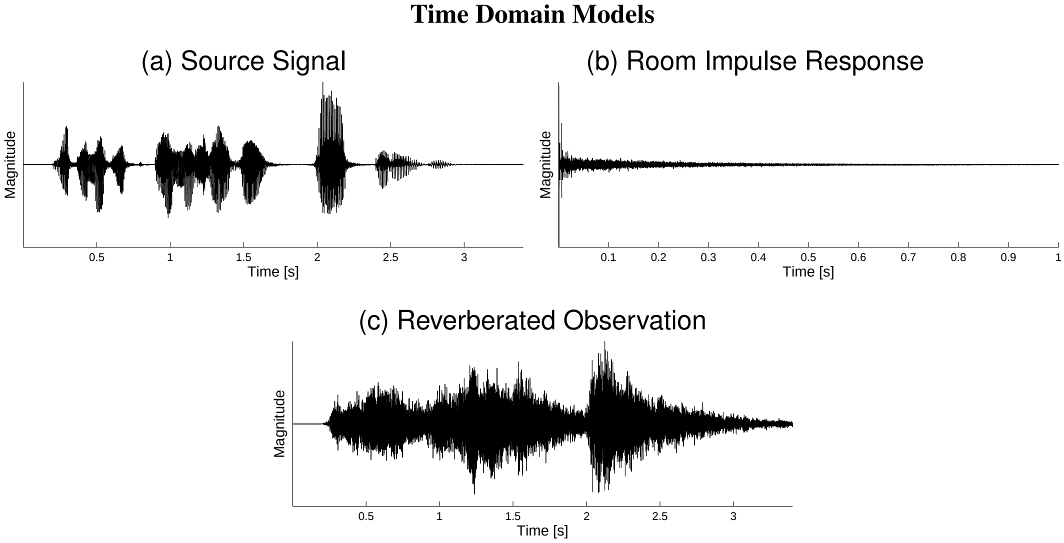

Figure 2.2: (a) Time domain model of the original signal. (b)Time domain model of the room impulse response. It can be observed that the impulse response has a decreasing nature over time. (c) Convolution result of the source signal and the room impulse response. No noise is added in order to show the effects of the reverberation.

The estimation in the thesis is obtained using convex optimization processes which is explained in the next section. However the optimization problems are formed using STFT coefficients. The trick required in order to define problems in STFT is explained in chapter 3.

The definition of the STFT is also given in the other chapters, because in each of the scenario different properties of the STFT are exploited. STFT is just a linear time frequency decomposition. Instead of STFT any other linear time frequency transformation can be used [5]. However with different transformations the representation of RIR in these domains is required to be calculated.

2.2 Convex Optimization Procedure

In order to estimate the original signal from reverberated data convex optimization procedure is used through the thesis. The convex optimization procedure can be generalized as,

$$\min_{\bar{x}} ||\bar{y} - \bar{H}\bar{x}||_2^2 + \sum_i \lambda_i p_i(\bar{x}) \qquad (2.4)$$

Where $\bar{x} = [x_1\ x_2\ \cdots x_N]^T$, $\bar{y} = [y_1\ y_2\ \cdots y_N]^T$ and $\bar{H}$ denotes the mixture matrix with corresponding RIR. The quadratic term given in Eq.(2.4) minimizes the error between the estimate and the observations. λ_i's denote the weighting factors that determine the importance on penalty functions. $p(\cdot)$'s denote the convex penalty functions. In this thesis, a variety of penalty functions are proposed for single channel and multi channel cases.

Through the thesis convex minimization problems are solved iteratively. Douglas Rachford Algorithm is used in chapter 4 and chapter 5. So the definition of the algorithm is required.

Douglas-Rachford Algorithm

In a convex minimization process the solution may not be easily computed due to the nature of penalty functions. With numerous penalty functions employed in the minimization process, the computation of gradient is not cost effective. In order to reduce the complexity, problem can be solved by dividing into less complex sub problems [6]. Using these sub problems iteratively the algorithm converges to the optimal point iteratively of the penalty function.

In DR Algorithm the penalty function is divided into two as,

$$\hat{x} = \arg\min_{x} \left(f(x) + g(x)\right) \tag{2.5}$$

Where $f(x) + g(x)$ form the penalty function of interest. In order to solve requiring more divisions, variable splitting can be used. In this formulation, the estimate $\hat{x}$ is found iteratively. For iterations, the concept 'proximity operator' is used. The proximal of a function of f is defined as,

$$J_{\gamma f}(x) = \arg\min_{a} \frac{1}{2\gamma}\|a - x\|_2^2 + f(a) \tag{2.6}$$

Using this knowledge the estimate can be calculated,

Algorithm 1 Douglas-Rachford Algorithm

```
$a^0 \leftarrow x$ , $\gamma \in (0,1)$, $\alpha \in \mathbb{R}^+$, $h \leftarrow 0$
repeat
    $b \leftarrow \left(2J_{\gamma f}(a^h) - a^h\right)$
    $a^{h+1} \leftarrow a^h(1-\alpha) + \alpha\left[2J_{\gamma g}(b) - b\right]$
    $h \leftarrow h + 1$
until convergence criterion met
$\hat{x} \leftarrow J_{\gamma f}(a^N)$
```

using this iterative algorithm. It can be observed that, the algorithm basically tries to satisfy the constraints one by one. The algorithm finds the optimum point that has the minimum distance to the contraint subsets. The convergence of this algorithm can be checked from [6].

The proximal definition is made in Eq.(2.6). Using this definition proximity operators of widely used functions can be given as examples.

Proximity of ℓ_1

The sparsity of a function is checked widely using the ℓ_1 norm. The penalty function of this norm is defined as,

$$p(x(n)) = \lambda \|x(n)\|_1 = \sum_n |x(n)| \tag{2.7}$$

The proximity of this penalty function is required in many applications, especially if used with other penalization functions, where taking gradient is not straightforward. In order to find the proximity operator, the minimization procedure is required to be solved. The problem,

$$J_{\gamma p}(x) = \arg\min_t \frac{1}{2\gamma}\|x-t\|_2^2 + \|t\|_1 \tag{2.8}$$

Here in the Eq.(2.8), both the quadratic term and the ℓ_1 term are convex. Therefore, it can be said that 0 is an element of the gradient of this function. If the gradient is taken with respect to t and it is assumed that $\hat{t}$ is the solution to the minimization problem in Eq.(2.8). The result,

$$0 \in (\hat{t} - x) + \gamma \mathrm{sgn}(\hat{t}) \tag{2.9}$$

Vector derivation can be checked from [7]. Here sgn denotes the sign function. This relation can be extended into two sub definitions as,

$$\begin{aligned} 0 &\in (\hat{t} - x) + \gamma \quad \text{where } \hat{t} > 0 \\ 0 &\in (\hat{t} - x) - \gamma \quad \text{where } \hat{t} < 0 \end{aligned} \tag{2.10}$$

The solution can be obtained with satisfying the equations and modifying the domain constraints on $\hat{t}$ with respect to x and λ. The solution can be obtained as,

$$\hat{t} = J_{\gamma p}(x) = \begin{cases} x - \gamma & \text{if } x - \gamma > 0 \\ x + \gamma & \text{if } x + \gamma < 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

The function denoted in Eq.(2.11) results in soft thresholding operator. It can be observed that with the ℓ_1 norm the magnitudes of the coefficients in the matrix are cropped.

This proximity operator definition also refers to ℓ_1 norm minimization problem, which is also used in chapter 3.

Proximity of ℓ_2

The ℓ_2 norm proximal calculation is taken into consideration. The ℓ_2 norm can be defined as,

$$p(x) = ||x(n)||_2 = \sqrt{\sum_n |x(n)|^2} \tag{2.12}$$

In order to find the proximity operator of this function, the minimization procedure can be used as,

$$J_{\gamma p}(x) = \arg\min_t \frac{1}{2\gamma}||x-t||_2^2 + \lambda\,||t||_2 \tag{2.13}$$

Where sgn denotes the sign function. In order to solve this problem the matrix notation for the ℓ_2 norm can be used. Assuming that $||x||_2 = \left(x^T x\right)^{\frac{1}{2}}$, the problem defined in Eq.(2.13) can be modified as,

$$\arg\min_t \frac{1}{2\gamma}(x-t)^T(x-t) + (x^T x)^{\frac{1}{2}} \tag{2.14}$$

Square of the ℓ_2 norm and ℓ_2 norm itself are convex. Therefore, in order to minimize the problem, the solution can be searched where 0 is the element of the gradient. Taking the gradient yields the result,

$$0 \in (\hat{t}-x) + \lambda\,(\hat{t}^T\hat{t})^{-\frac{1}{2}}\hat{t} \tag{2.15}$$

From this point it can be observed that $\hat{t} = cx$ where c is a scaling factor. Thus, the proximal calculation defined in Eq.(2.13), can be modified into the problem,

$$\begin{aligned} &\min_c \frac{1}{2}||cy-y||_2^2 + \gamma c||y||_2 \\ =&\min_{c\geq 0} \frac{1}{2}(c-1)^2||y||_2^2 + \gamma c||y||_2 \\ =&\min_{c\geq 0} \frac{1}{2}(c-1)^2 + \gamma\frac{c}{||y||_2} \end{aligned} \tag{2.16}$$

If the gradient with respect to c is set the zero, it can be found $c = 1 - \gamma/||y||_2$. Therefore, the solution can be obtained as,

$$\hat{t} = J_{\gamma p}(x) = \left(1 - \frac{\gamma}{||y||_2}\right)y \tag{2.17}$$

It can be observed that the proximity operator of the euclidean norm thresholds the magnitude of the given vector. Proximity on ℓ_2 norm is used in chapter 5 for calculating proximity on mixed norm.

3. A DEREVERBERATION FORMULATION BASED ON SPARSITY

The dereverberation problem can be cast as a sparsity based minimization with a known room impulse response. By making use of sparse nature of the time-frequency transform coefficients of the original signal estimate can be enhanced. Short time Fourier transform is used as the time frequency decomposition. In such scenarios where the room impulse response is longer than the window employed in STFT, the convolution operator can not be described as an element wise multiplication in STFT domain. This concept is explained in the chapter 2. Room impulse response, which is a convolutive linear operator, is defined on time domain and the sparsity is defined on time-frequency spectrum. As the conditions are defined in different domains, transform operator and its inverse are required in each iteration. The computation cost of the algorithm increases dramatically, because time required to change domains is a burden. This problem can be prevented if the room impulse response is modeled in the transform domain. Besides, sparsity constraint is set on the transform domain, it can not be modeled in time domain, which makes the STFT domain definition of RIR mandatory.

With the room impulse response expression in transform domain, the dereverberation problem can be posed as a convex minimization problem on a frequency band of time frequency transform spectrum with sparsity constraint. Therefore, the iterations of this problem are free of the transform which is expected to increase computational efficiency. The dereverbed signal can be obtained after these iterations with taking the inverse transform only once.

In this chapter sparsity based dereverberation problem is cast using only time frequency coefficients. For that purpose effects of the RIR is modeled in STFT. This model proposes a filter for each frequency band of the transform coefficients, when convolved with the corresponding band, represents the effects of the reverberation. Therefore, this results in decreased computational burden and allows usage of other penalty functions in STFT domain. This concept is also exploited in other chapters in order to use penalty functions defined on the time frequency spectrum. In addition, dereverberation with sparsity constraint is obtained in this chapter. Both the RIR estimates and the derevereberation algorithm are justified in the experiments section.

3.1 Proposed Method

In this section a methodology for estimating the original signal from the reverberated and noisy observation is proposed. The single channel observations in time domain can be modeled as,

$$y = Hx + n \tag{3.1}$$

Where x denotes the source signal, H is the convolution operator with the room impulse response h, n is white Gaussian noise and y is the observation.

The convex minimization procedure is posed using a penalty function, this penalty function is formed by using a quadratic term, which penalizes the difference from the observation and an ℓ_1 term which enforces sparsity of the spectrogram. With assumption, S denotes the STFT operator and S^* denote the adjoint operator. STFT forms a tight frame, thus it can be said that $S^*S = I$. The problem is defined as,

$$\min_X \frac{1}{2}\|y - HS^*X\|_2^2 + \lambda\|X\|_1 \tag{3.2}$$

This notation is defined in [8] for sparse dereverberation. Here X denotes the STFT of x. This problem can be solved using Iterative Shrinkage Thresholding Algorithm (ISTA) which has the iterations as,

$$\begin{aligned} &\hat{X}^{k+1} = T_{\alpha\lambda}\left(\hat{X}^k + \alpha\,(HS^*)^*\left(y - HS^*\hat{X}^k\right)\right) \\ &\text{where } T_\alpha(x) = \begin{cases} 0, & \text{if } |x| < \alpha \\ \frac{1-\alpha}{|x|}x, & \text{if } |x| \geq \alpha \end{cases} \end{aligned} \tag{3.3}$$

It can be easily observed that during the iterations the domain is changed twice. In order to enforce sparsity of the STFT coefficients using thresholding the transform is required. After this step deconvolution with RIR is required. Deconvolution is performed using time domain representation of the signal. Therefore, an inverse transform is required. Computational time required for changing to STFT domain from time domain is considerably high compared to thresholding operator. As in thresholding elementwise multiplication and addition is required where within STFT an elementwise multiplication is done and FT is taken over a windowed portion.

In order to prevent domain changes in the solution, an operator is defined. The operator $\mathscr{H}$ represents the effects of the reverberation in STFT domain as,

$$SH \approx \mathscr{H} S \tag{3.4}$$

Here $\mathscr{H}$ is the operator that represents the convolution of the room impulse response in STFT domain as explained in [9]. The definition of this operator is given in the next section.

Using this operator, a new problem can be defined in STFT domain as,

$$\hat{X} = \arg\min_{X} \frac{1}{2} \|Y - \mathscr{H} X\|_2^2 + \lambda \|X\|_1 \tag{3.5}$$

It can be observed that the problems in (3.2) and (3.5) are not equal but the results are equivalent. Both problems check the sparsity of the spectrum and the squared error between the estimate and the observation with known RIR.

This problem can again be solved using ISTA with the iterations,

$$\hat{X}^{k+1} = T_{\alpha\lambda}\left(\hat{X}^k + \alpha \mathscr{H}^* \left(Y - \mathscr{H}\hat{X}\right)\right) \tag{3.6}$$

Since the iterations does not include domain changes (S or S^*), iterating using $\mathscr{H}$ is more beneficial compared to calculating HS^*. The aim is to find such $\mathscr{H}$ that approximately satisfies the reverberation effects in STFT domain. This idea is used for single channel observation in [9] and for multichannel case in [10], however in both applications sparsity is neglected.

An operator $\mathscr{H}$ can be found if Short Time Fourier Transform windowing function is longer than the impulse response. Under this condition the operator directly represents elementwise multiplication with the STFT of the RIR in the STFT domain. However, typical windowing functions are in 30[ms] - 60[ms] interval where impulse responses are around few hundreds of [ms]. As explained it is not possible to find a perfectly fitting operator $\mathscr{H}$, the aim is to estimate it by solving a linear system. The next section explains the basics of the operator.

3.1.1 Room impulse response estimate in STFT

In order to determine a filter that represents the effects of reverberation operator, the properties of STFT should be exploited. In order to find such a filter definition of an STFT frequency is needed. In this section the filter $\mathscr{H}$ is explained [9].

The filter is defined exploiting the filter bank representation of the STFT where this relation is also visualized in Fig.3.1. Let $(\downarrow N)(\cdot)$ denote the down sampling with N, $*$ denote the convolution and under the assumption that $g(n)$ is a low pass filter, STFT can be defined as,

$$\begin{aligned} g_k(n) &= g(n)\exp(-jk\Delta_w n) \\ x_k(n) &= (\downarrow N)(x(n) * g_k(n)) \end{aligned} \tag{3.7}$$

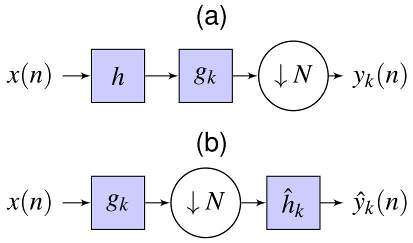

Figure 3.1: (a) Filter bank representation of one STFT frequency band after a filter h. (b) Representing the effects of the convolution on STFT coefficients.

$x_k(n)$ denotes the k^{th} frequency band of the STFT. The aim is to find a filter $\hat{h}_k(n)$ for each k^{th} frequency band that satisfies,

$$x_k(n) * \hat{h}_k \approx (\downarrow N)((x(n) * h(n)) * g_k(n)) = y_k(n) \tag{3.8}$$

this relation.

Different from previous notation, where the capital letters denote the STFT coefficients, FT of the frequency bands are denoted with capital letters for this section. For example the transform of $g_k(n)$ is denoted in the form $G_k(w)$. This notation can be distinguished from STFT coefficients as the input variable of the function is denoted with w. Assuming that the windowing function used in STFT is band limited and s_k is the center frequency, then,

$$G_k(w) = 0 \text{ if } w \in \left[s_k - \pi, s_k + \frac{\pi}{N}\right] \cup \left[s_k + \frac{\pi}{N}, s_k + \pi\right] \tag{3.9}$$

can be written. This says the filters G_ks are band limited.

Therefore, the relation expressed in Eq.(3.8) can be written in Fourier domain, assuming that filters are band limited the relations are defined as,

$$Y_k(w) = X\left(\frac{w}{N}\right) H\left(\frac{w}{N}\right) G_k\left(\frac{w}{N}\right) \text{ if } w \in [Ns_k - \pi, Ns_k + \pi] \tag{3.10}$$

$$Y_k(w) = X\left(\frac{w}{N}\right) \hat{\mathscr{H}}(w) G_k\left(\frac{w}{N}\right) \text{ if } w \in [Ns_k - \pi, Ns_k + \pi] \tag{3.11}$$

As the intention is to find the filter that satisfies Eq.(3.8), the solution can be formed combining Eq.(3.10) and Eq.(3.11) yields,

$$\hat{\mathscr{H}_k}(w) = H\left(\frac{w}{N}\right) \text{ if } w \in [Ns_k - \pi, Ns_k + \pi] \tag{3.12}$$

But in practice it is not possible to find a perfectly band limited window as defined in (3.9). Then, it can be observed that the approximate $\hat{h}_k$s have a different effect compared to h. In order to represent the effects of the RIR, it is desired to minimize the squared error between the linear systems defined in Eq.3.10 and Eq.(3.11) can be minimized. The filter can be found through the minimization process,

$$\hat{\mathscr{H}_k} = \arg\min_U \int_{w=Ns-\pi}^{Ns+\pi} \left\| G_k\left(\frac{w}{N}\right) \left[H\left(\frac{w}{N}\right) - U(w) \right] \right\|_2^2 dw \tag{3.13}$$

Thus $\hat{\mathscr{H}_k}$s are the optimum filters that represent the effects of the RIR in corresponding frequency band.

The minimization given in Eq.(3.13) is calculated for all k. Calculated filters $\hat{h}_k$s represent the effect of the room impulse response in the corresponding channel. Thus it is known that with the filter one can obtain,

$$Y_k(n) = X_k(n) * \hat{h}_k(n) \tag{3.14}$$

Using this relation, responses for corresponding channels can be computed. Using these filters one can use the estimate in STFT coefficients. Assume that $\mathscr{H}$ is the operator that maps STFT coefficients of x onto STFT coefficients of Y. The relation is;

$$\hat{Y} = \mathscr{H} Sx \approx SHx = Y \tag{3.15}$$

Solution $\mathscr{H}$ denotes the operator that applies the effect of RIR in time frequency spectrum (computing Eq.(3.14) for all frequency bands k). This notation shows that convolving each frequency band with corresponding estimate, approximately results in the coefficients of the reverberated observation. It can be also noted that $\mathscr{H}_k$ represents convolution with the IFT of the calculated optimal filter $\hat{\mathscr{H}_k}$

In order to form a minimization procedure for a specified frequency band let $\mathscr{H}_k$ denote the estimate of RIR for k^{th} channel as defined before. In the section the minimization procedure is presented.

3.1.2 Estimation on STFT coefficients

Using the filter estimate given in the previous section one can form a new observation model instead of Eq.(3.1). The problem can be modeled for each frequency band as,

$$Y_k = \mathscr{H}_k X_k + \tilde{U}_k \text{ where } \tilde{U}_k \text{ is the channel noise} \tag{3.16}$$

Here $\mathscr{H}_k$ denotes the convolution operator with the estimated impulse response in STFT from Eq.(3.13).

In the equation $\tilde{U}_k$ does not only represent the effects of the Gaussian noise n given in Eq.(3.1) but also represents the errors caused by the room impulse response estimate. The advantage of this formulation, is one can penalize the sparsity of the frequency band. Then, using this notation, a minimization problem can be formed using only time frequency coefficients. A specific frequency band can be estimated using the problem,

$$\hat{X}_k(n) = \arg\min_z \frac{1}{2}||Y_k(n) - \mathscr{H}_k z(n)||_2^2 + \lambda ||z(n)||_1 \tag{3.17}$$

in this form. In Eq.(3.17) $X_k(n)$ denotes the n^{th} time bin of the k^{th} frequency band of the STFT coefficients. This problem indicates that each channel is treated separately. This problem can be solved using ISTA again. Assuming that $\mathscr{H}_k^*$ represent the conjugate of the convolution operator $\mathscr{H}_k$. Thus, this denotes the convolution with time reversed conjugate of the original filter. ISTA can be posed as,

Algorithm 2 Iterative Shrinkage Thresholding Algorithm

```
1: repeat
2:     $\hat{X}_k(n) \leftarrow \hat{X}_k(n) + \alpha \left( \mathscr{H}_k^* \left( Y_k - \mathscr{H}_k \hat{X}_k \right) \right), \quad \forall k$
3:     $\hat{X}_k(n) \leftarrow T_{\lambda\alpha}\left( \hat{X}_k(n) \right), \quad \forall k, n$
4: until convergence criterion met
```

This algorithm converges if α is chosen small enough.

Consider σ_k to be the biggest eigenvalue of $\mathscr{H}^*\mathscr{H}_k$. If $\alpha\sigma_k < 2$ is satisfied it is guaranteed that the algorithm converges to a solution of the penalty function given in Eq.(3.17) [10].

The operator $T_{\lambda\alpha}$ is defined previously in Eq.(3.3) as the soft thresholding operator. ISTA can be placed into the overall algorithm for the estimation process. The overall algorithm can be posed as,

Algorithm 3 Dereverberation with ISTA

```
$\alpha$ from [10], $\lambda \in \mathbb{R}^+$
$\hat{X} \leftarrow 0$
repeat $\forall k$
    $\mathscr{H}_k$ from $H_k$                                   Eq.(3.13)
    repeat
        $\hat{X}_k \leftarrow \hat{X}_k + \alpha \left(\mathscr{H}_k^* \left(Y_k - \mathscr{H}_k \hat{X}_k\right)\right)$
        $\hat{X}_k \leftarrow T_{\lambda\alpha}\left(\hat{X}_k\right)$,
    until convergence criterion met
until finished
$\hat{x} \leftarrow S^* \hat{X}$
```

Using this formulation dereverberation in STFT can be achieved. In next chapter examples are demonstrated using this algorithm.

3.2 Experiments and Discussion

In order to justify the performance of room impulse response estimate and the algorithm for dereverberation, a series of experiments are performed.

A measured room impulse response of length 1[s] is used with sampling frequency 44.1[kHz]. This impulse response is converted into frequency band filters that poses the effects of RIR for the corresponding frequency band. The conversion is justified in Fig.3.2. In order to form the figure an active band from the original signal, coefficients are reverberated using the filter estimate. The same corresponding band is chosen from the STFT of conventionally reverberated signal in time domain and comparison is performed. In Fig.3.2-(a) it can be seen that the imaginary parts of the time domain reverberated signal and the estimate channels fit. From Fig.3.2-(b), it can be observed that the difference between the absolute values is negligible. Thus, it is proven that using the modified room impulse response for STFT domain is suitable, as the difference caused by RIR estimates are negligible.

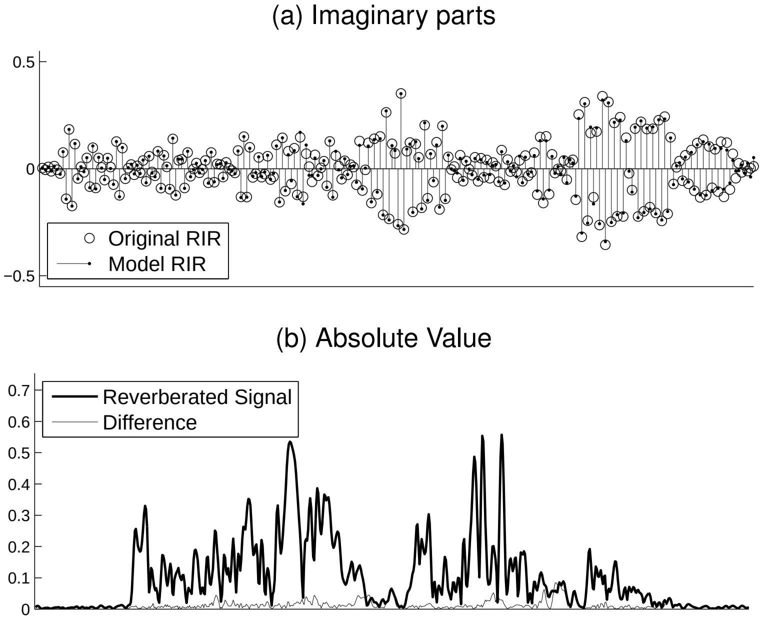

Figure 3.2: For this experiment an active channel of the STFT is selected. (a) Imaginary parts of the original and model coefficients. (b) Absolute value of the band coefficients.

Through experiments a 3[s] long speech signal is used with 44.1[kHz] sampling frequency. This signal is synthetically reverberated using the measured room impulse response and noise is added with predetermined SNR value 10[dB]. The original signal can be checked from Fig.3.3-(a). The reverberated and noisy signal can be checked from Fig.3.3-(b). In this figure the effects of reverberation can be observed. The harmonics of the original signal are extended through consecutive time coefficient in the corresponding band. It can be said that this effect can be observed because length of room impulse response is longer compared to the STFT window. Otherwise reverberation can be modeled as an element wise multiplication of the STFT coefficients in the transform domain, leaving an observation harmonic structure at the same length of the source signal. In addition, the effects of the noise can be observed as the activity in the time frequency coefficients outside the harmonics increased.

The estimate using proposed method is visualized in Fig.3.3-(c). It can be observed that, in the estimate harmonics are shorter compared to the observation and similar to the source signal. Therefore one can say that dereverberation is achieved. Also it can be stated that denoising is achieved, as the spectrum is sparse. It can be observed that,

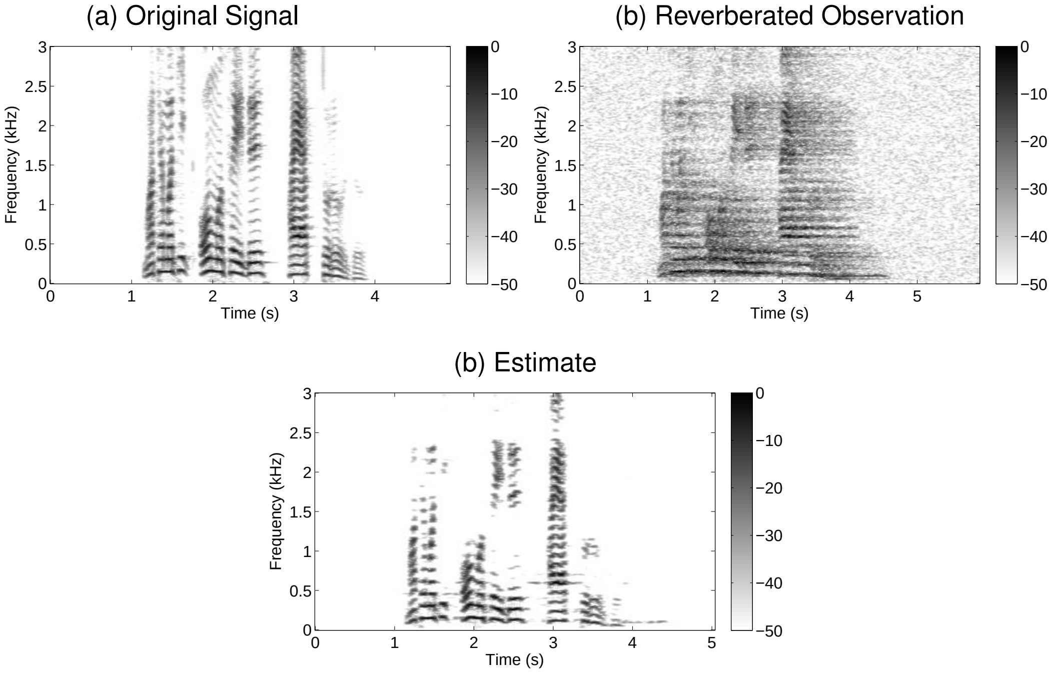

Figure 3.3: For this experiment an active channel of the STFT is selected. (a) Imaginary parts of the original and model coefficients. (b) Absolute value of the band coefficients.

coefficients with low magnitudes are suppressed. However during this process some of the harmonics are lost. This can be explained with the magnitudes of the harmonics being close to 0. With the increasing presence of noise these harmonics tend to lose their information.

In order to visualize the reconstruction an active frequency band of the spectrogram can be checked. In Fig.3.4 STFT coefficients are compared using both real parts and imaginary parts. Figures show that the estimation in this active frequency band is also fitting.

The estimation is also questioned with different input SNR values. Assuming that input SNR is a preset value between the reverberated observation and the noisy reverberated observation. As expected, estimate SNR values increase as lambda increases for worse input SNR values. As noise increases, the observation carries less information about the original signal. This can be clarified as the presence of the noise increases, sparsity constraint becomes more reliable. This is visualized in Fig.3.4. It is clear that as input SNR increases the estimation reaches its maximum value at lower λ values. It can also be observed from Eq.(3.17) $\lambda = 0$ case is equal to the LSE. The estimate where

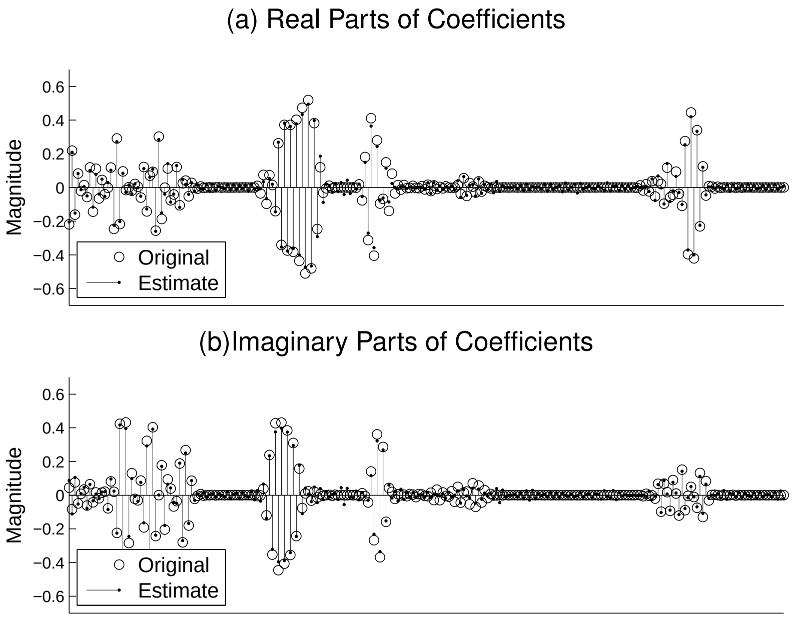

Figure 3.4: The estimate and source signal coefficients for an active frequency band. (a) Real parts of the coefficients. (b) Imaginary parts of the coefficients.

$\lambda \neq 0$ yields better SNR values. Therefore, the proposd method yields better results compared to LSE.

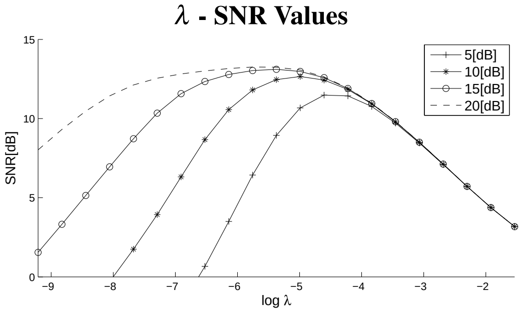

Figure 3.5: In this figure it is justified that the method requires higher values ofλ in order to achieve their best SNR values. Also the input SNR values determine the quality of the estimate as well. The input SNR values are higher compared to maximum achieved values. The reason is, estimate and the observation SNR values are calculated differently.

Overall, it can be observed from the experiments that the dereverberation algorithm shortens the effects of room impulse response and reduces the effects of noise. Also the STFT and its inverse is note used in the algorithm decreases computational time. Through the experients it is also noticed that the computational time required for the time domain derevereberation is halved.

4. DEREVERBERATION WITH EMPLOYING PHASE INFORMATION

In that chapter dereverberation problem is solved employing the phase information. In time domain observations are modeled,

$$y = Hx + n \tag{4.1}$$

in that form. It can be observed that single channel system model is similar to definition in chapter 3. In the equation Eq.4.1 y denotes the noisy and reverberated observation, x denotes the original signal, H is the convolution operator with the known impulse response h, n is white Gaussian noise.

The dereverberation is performed using sparse nature of STFT coefficients, using ℓ_1 type penalty function as proposed in chapter 3. But, with the effects of reverberation, STFT coefficients diverge from being sparse. Also a solution with modifying only the magnitudes of the STFT coefficients yields a solution with musical noise. Nevertheless, even with these outcomes the sparse nature of the time frequency spectrum of an audio signal is robust to noise. This is why, magnitude information can not be neglected and used in conventional methods. However, in order to compensate the musical noise phase information can be taken into consideration. Phase information is effected by noise dramatically, also fragile to the effects of reverberation. However, even with these effects, phase information can be used to increase the efficiency if employed. In that chapter, a method that employs magnitude and phase information is given.

In a harmonic of a specified frequency band of STFT of a signal, time consecutive coefficients have a correlated phase information. This relation can be referred as a constant phase shift between coefficients. Phase difference between two STFT coefficients on any harmonic is assumed to be approximately equal. Considering each audio signal as a linear combination of sinusoidals, it is assumed that, there lies a complex exponential mapping one coefficient to another one in the harmonic [11–13]. Outside the harmonics magnitudes of the coefficients are 0 due to no activity, which makes the phase information meaningless. The phase within the harmonics can be linked to other coefficients with a complex exponential and this property can be disregarded outside the harmonics. So generalizing the property for a frequency band is feasible. Thus, considering that a coefficient can be represented with the time consecutive coefficient, a frequency band of the transform can be linked with phase shifted version of itself. As defined before a complex number can be found for each harmonic in the frequency band. Also it is known that there are few harmonics in a frequency band. Thus a piece wise constant mask can be found for each frequency band. The mask takes 0 value and a non-zero constant through each harmonic. outside the harmonics. The mask can be obtained after a minimization procedure, which is solved iteratively. Using that procedure, both phase shift stays constant with same magnitude and sparsity in a frequency band is satisfied. Therefore, the harmonic structure is preserved while suppressing the noise coefficients. This mask is applied on the LSE of the signal as it requires an initial time frequency spectra to enhance.

In order to remark it is also assumed that a convolutive impulse response operator in the STFT domain can be found. Where the problem in Eq.(4.1) can be defined in STFT,

$$Y = \mathscr{H}X + U \tag{4.2}$$

in that form. Here the operator $\mathscr{H}$ is the convolutive reverberation operation in STFT. If the k^{th} row of the operator is convolved with the k^{th} frequency band of the STFT coefficients it yields the k^{th} frequency band of the observation. Assume that the STFT operator is denoted with S where $X = Sx$. Assume that the k^{th} band of the coefficients is denoted with sub indent $\cdot_k$ and the convolution operation is denoted with $*$. Then the relation between $\mathscr{H}$ and H can be shown,

$$Y_k = [S(Hx)]_k = X_k * \mathscr{H}_k \tag{4.3}$$

in this form [9]. This form is defined in the previous chapter . The definition is assumed to be satisfied for this chapter as well.

4.1 Proposed Method

In that section, a method for dereverberation is proposed. In order to employ phase information for that process, the phase relation between coefficients is explained.

Assume that k^{th} frequency band of the signal is taken into consideration. In that frequency band, it can be observed that the activity is concentrated in the harmonics and the remaining coefficients are zero. The coefficients in the harmonics tend to have similar magnitude in STFT spectrum. Because each harmonic represented in the spectrum is considerably short for an audio source to change its magnitude dramatically. These coefficients can be represented as a vector with constant phase shift in the complex plane. The equivalent magnitude and constant phase shift relation is posed in [14]. This property can be modeled with a relation,

$$\frac{X(k,l+1)}{X(k,l)} \approx \frac{X(k,l+n+1)}{X(k,l+n)} \tag{4.4}$$

in this form. The relation can also be explained with the model posing a complex exponential between each coefficient in the harmonic Therefore, Eq.(4.4) sates that there is only one complex exponential mapping consecutive coefficients in a harmonic. This relation can be also verified using sinusoidal models, where the audio signal can be represented as a superposition of sinusoidals [11]. In the STFT spectrum of an audio signal it is observable that the complex exponential does not satisfy the relation between coefficients and instead using a complex number can form a robust method to map coefficients. The relation can be shown,

$$X(k,l) \approx X(k,l+1)\alpha_k(l) \tag{4.5}$$

in this form. This relation however becomes meaningless outside of the harmonics. As explained before, it is known that outside the harmonics there is no activity and with magnitudes approaching 0 phase information becomes meaningless. As the phase shift assumed to be constant, the complex vector is expected to be constant. A penalty function in order to force the phase shift to be constant, can be formed summing the phase differences between the coefficients regardless of convexity. However, the penalty function is desired to be convex. Thus, the mask defined in Eq.(4.5) is the point of escape.

In order to form a convex penalty function, the phase shifted version of the signal is defined,

$$\tilde{X}(k,l) = \frac{|X(k,l)|}{|X(k,l+1)|} X(k,l+1) \tag{4.6}$$

in that form. Using the constancy relation in phase difference denoted in Eq.(4.4), the phase shifted version and the signal can be linked,

$$X(k,l) \approx \tilde{X}(k,l)\alpha_k(l) \tag{4.7}$$

using this relation. Here α is a complex valued vector. This vector should also satisfy the properties of the mask proposed in Eq.(4.5). Thus α is required to be piecewise constant [14]. In coefficients harmonics are assumed to have a constant shift, this leads to constancy which can be explained linking the relation with Eq.(4.7) and harmonics also have a sparse nature, which is a property of time frequency spectrum.

In that scenario the room impulse response and the properties of the noise are assumed to be known. Thus, in the proposed method, the initial point can be set to the LSE. Assume that $\mathscr{H}$ denotes the room impulse response operator in STFT domain and σ^2 denotes the variance of the noise. The least squares estimate (LSE) can be obtained by solving the complex minimization problem,

$$X_{LSE} = \arg\min_{x} E\left\{\|Y - (\mathscr{H}X + U)\|_2^2\right\} \tag{4.8}$$

of that form. LSE only includes a quadratic term which is convex. Therefore, problem has the 0 in its gradient. The solution to that problem can be found by taking the gradient and setting it to zero. LSE is,

$$X_{LSE} = (\mathscr{H}^H Y)/\left(\|\mathscr{H}\|_2^2 + \sigma^2\right) \tag{4.9}$$

calculated in that form. Estimating the original signal can be achieved by masking the phase shifted version X_{LSE}. Let $\tilde{X}_{LSE}$ denote the phase shifted version,

$$\tilde{X}_{LSE}(k,l) = \frac{|X_{LSE}(k,l)|}{|X_{LSE}(k,l+1)|} X(k,l+1) \tag{4.10}$$

as given. Let $\hat{X}$ denote the estimate, which can be obtained masking the phase shifted version. This masking is proposed exploiting the mapping between two consecutive time coefficients. The masking can be shown,

$$\hat{X}(k,l) = \alpha(k,l)\tilde{X}_{LSE}(k,l) \tag{4.11}$$

in that form. It can be observed that, the difference between $\hat{X}$ and $\tilde{X}_{LSE}$ is modeled with a mask. That mask is the same as the vector defined in Eq.(4.7). This relation also preserves the harmonic structure. With masking using the result of the minimization procedure, quality of the LSE is expected to be increased. The mask can be obtained using a minimization process, enforcing it to be piecewise constant.

4.1.1 Minimization on mask

The optimum mask for estimate can be obtained using a convex optimization process. In order to form the mask α, the vectors for each frequency band is calculated. The vectors are required to be piece wise constant, as explained previously. The minimization problem can be posed,

$$\hat{\alpha}_k = \arg\min_{\alpha} \frac{1}{2}\|Y - \mathscr{H} X_{LSE}\alpha_k\|_2^2 + \lambda_1 \mathrm{TV}(\alpha_k) + \lambda_2\|\alpha_k\|_1 \tag{4.12}$$

in that form. Here ℓ_1 enforces the sparsity. The mask is here calculated to be applied on the phase shifted version of the LSE. The relation of the sparse nature of the phase shifted version and the original signal is analogous. Sparsity is a property that takes magnitudes into consideration. Shifting phase does not change the sparse nature of the signal. Thus, this is proper to modify the phase shifted version with a sparse mask. This can also be justified with the usage of the sparse mask on the phase shifted version.

$\mathrm{TV}(\cdot)$ enforces the constancy of the mask. The operator can be defined,

$$\begin{gathered}\mathrm{TV}(v_k) = \|Dv_k\|_1 \\ D = \begin{bmatrix} 1 & -1 & & & 0 \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ 0 & & & 1 & -1 \end{bmatrix}\end{gathered} \tag{4.13}$$

in that form. Observe that the TV norm expressed in Eq.(4.12) is applied on each frequency band. Thus, it is noticeable that the TV norm penalizes the differences between consecutive elements in the vector. In the minimization problem defined in Eq.(4.12), TV norm penalizes the magnitude difference of the consecutive coefficients in the vector.

The minimization problem given in Eq.(4.12) can be solved for each time frequency component in order to form the mask. The proposed method can be compared to non-negative garrote in [15]. In non-negative garrote the problem is generalized with a sparsity constraint. In the proposed method, the constancy of the mask is also taken into consideration. Thus, the phase information is used to preserve harmonic structure.

In the next section a method is given in order to solve the proposed problem in Eq.(4.12).

4.1.2 Solution of the minimization

The problem proposed in Eq.(4.12) is a convex minimization problem. The problem cannot be solved directly calculating the gradient, due to the complexity caused by the TV norm and the ℓ_1. However, it can be solved using the DR algorithm variable splitting [16]. DR is explained in chapter 2 Alg.1. The next section clarifies the variable splitting for DR Algorithm. For simplicity, instead of α_k, α is used as the vector of interest in the corresponding frequency band.

4.1.3 Variable splitting

In this section, a variable splitting for Eq.(4.12) is proposed. With a penalty function including total variation norm, variable splitting can be achieved with the definition of two new variables $z = \alpha$ and $u = D\alpha$ [17]. In order to preserve the original problem, while using newly defined variables, characteristic function that links these variables with each other is defined,

$$i_c(\alpha, u, z) = \begin{cases} 0, & \text{if } u = D\alpha \text{ and } z = u \\ \infty & \text{otherwise} \end{cases} \tag{4.14}$$

in that form. The relationship between newly defined variables is enforced using the characteristic function. It can be observed that the characteristic function can not be minimized if the conditions are not met, as it takes the infinity value. The function gets zero only if the conditions are met. If the characteristic function is inserted into the original problem with new variables into Eq.(4.12), the new penalty function,

$$\begin{aligned} [\hat{\alpha}, \hat{u}, \hat{z}]^T &= \arg\min_{\alpha,u,z} f(\alpha, u, z) + g(\alpha, u, z) \\ f(\alpha, u, z) &= \frac{1}{2}||Y - \mathscr{H} X_{LSE}\alpha||_2^2 + \lambda_1 ||u||_1 + \lambda_2 ||z||_1 \\ g(\alpha, u, z) &= i_c(\alpha, u, z) \end{aligned} \tag{4.15}$$

can be obtained in this form. The problem defined in Eq.(4.15) is equal to the problem defined in Eq.(4.12). In order to apply DR Algorithm the proximals are required. Proximal values are calculated in the next section,

Proximal Calculation

The division of the penalty function is given in Eq.(4.15). Proximals for each of these functions are required for the algorithm. Proximal for the function f can be calculated,

$$\begin{aligned} J_{\gamma f}(\alpha,u,z) = [\tilde{\alpha}\ \tilde{u}\ \tilde{z}]^T = & \arg\min_{a,b,c} \frac{1}{2\gamma}\left[\|\alpha-a\|_2^2+\|u-b\|_2^2+\|c-z\|_2^2\right]\cdots \\ & +\frac{1}{2}\|Y-\mathscr{H}X_{LSE}a\|_2^2+\lambda_1\|b\|_1+\lambda_2\|c\|_1 \end{aligned} \tag{4.16}$$

in that form.

It can be easily observed that function is separable with respect to the variables,

$$\begin{aligned} \tilde{\alpha} &= \arg\min_{a} \frac{1}{2\gamma}\|\alpha-a\|_2^2+\frac{1}{2}\|Y-\mathscr{H}X_{LSE}a\|_2^2 \\ \tilde{u} &= \arg\min_{b} \frac{1}{2\gamma}\|u-b\|_2^2+\lambda_1\|u\|_1 \\ \tilde{z} &= \arg\min_{c} \frac{1}{2\gamma}\|z-c\|_2^2+\lambda_2\|z\|_1 \end{aligned} \tag{4.17}$$

in that form. All the functions that are going to be minimized are convex. Thus, the proximal values for each variable can be found by setting the gradient to 0. Here, in order to preserve simplicity, $Q=\mathscr{H}\tilde{Y}$ is defined. The results,

$$\begin{aligned} \tilde{\alpha} &= \left(I+\gamma Q^H Q\right)^{-1}\left(\alpha+\gamma Q^H Y\right) \\ \tilde{u} &= T_{\lambda_1\gamma}(u) \\ \tilde{z} &= T_{\lambda_2\gamma}(z) \end{aligned} \tag{4.18}$$

are in that form. $T_{(\cdot)}(\cdot)$ denotes the soft thresholding operator defined in Eq.(3.3).

The proximal of g can be calculated using,

$$\begin{aligned} J_{\gamma g}(\alpha,u,z) = [\dot{\alpha}\ \dot{u}\ \dot{z}]^T = & \arg\min_{a,b,c} \frac{1}{2\gamma}\left[\|\alpha-a\|_2^2+\|u-b\|_2^2+\|v-z\|_2^2\right]\cdots \\ & +i_c(\alpha,u,z) \end{aligned} \tag{4.19}$$

that form. This function, however, is not separable with respect to variables, because the characteristic function links all the variables together. Because of the characteristic function, the penalty function g can only be minimized if the equality is satisfied (when the characteristic function gets the 0 value). This leads to the proximal functions for each variable,

$$\begin{aligned} \dot{\alpha} &= \arg\min_{a} \frac{1}{2}\left[\|\alpha-a\|_2^2+\|u-Da\|_2^2+\|v-a\|_2^2\right] \\ \dot{u} &= \arg\min_{b} \frac{1}{2}\left[\|\alpha-D^Tb\|_2^2+\|u-b\|_2^2+\|v-D^Tb\|_2^2\right] \\ \dot{z} &= \arg\min_{c} \frac{1}{2}\left[\|\alpha-c\|_2^2+\|u-Dc\|_2^2+\|v-c\|_2^2\right] \end{aligned} \tag{4.20}$$

in that form. Here D^T denotes the transpose of the operator D.

The proximals can be calculated taking the gradient and setting it to 0. The results can be obtained,

$$\begin{aligned}\dot{\alpha} &= \left(D^T D+2I\right)^{-1}\left(\alpha+D^T u+z\right)\\ \dot{u} &= D\left(D^T D+2I\right)^{-1}\left(\alpha+D^T u+z\right)\\ \dot{z} &= \left(D^T D+2I\right)^{-1}\left(\alpha+D^T u+z\right)\end{aligned}\tag{4.21}$$

in that form. It can be observed that γ is not important for the characteristic function proximal. The reason is that, the only case interested in the characteristic function is when it is equal to 0.

As the main ingredients of DR Algorithm are obtained, the algorithm can be formed. In the next section, the dereverberation algorithm employing phase information is explained.

4.2 Dereverberation Algorithm

Dereverberation is achieved by masking the phase shifted LSE. The LSE is estimated basically under the assumption that the room impulse response and the noise properties are known in Eq.(4.9). The phase of the LSE is shifted using Eq.(4.6). Then the algorithm calculates optimal mask vectors for each frequency band. The STFT coefficients of each frequency band for the mask are calculated for a convergence criterion. This criterion can either be a maximum number of iterations or an upper bound of tolerance. After obtaining the mask, it is used to obtain the estimate STFT coefficients. Taking the inverse STFT after masking yields the estimate. The algorithm can be posed,

Algorithm 4 Dereverberation Algorithm

```
$\gamma \in (0,1)$ , $\lambda_1 \in \mathbb{R}^+$ , $\lambda_2 \in \mathbb{R}^+$
$X_{LSE} = \left(\mathscr{H}^H Y\right) / \left(\|\mathscr{H}\|_2^2 + \sigma^2\right)$
$\tilde{Y} \leftarrow$ PhaseShift $(X_{LSE})$                                  Eq. (4.6)
**repeat** $\forall k \in \{0, K\}$,
    $\hat{\alpha}_k^0 \leftarrow \tilde{Y}_k$ , $\hat{u}_k^0 \leftarrow \hat{\alpha}_k^0$ , $\hat{z}^0 \leftarrow D\hat{\alpha}_k^0$, $h \leftarrow 0$
    **repeat**
        $A \leftarrow 2J_{\gamma f}(\hat{\alpha}_k^h, \hat{u}_k^h, \hat{z}_k^h) - \left[\hat{\alpha}_k^h \; \hat{u}_k^h \; \hat{z}_k^h\right]^T$          Eq. (4.18)
        $B \leftarrow 2J_{\gamma g}(A) - A$                                    Eq. (4.21)
        $\left[\hat{\alpha}_k^{h+1} \; \hat{u}_k^{h+1} \; \hat{z}_k^{h+1}\right]^T \leftarrow (1-\gamma)\left[\hat{\alpha}_k^h \; \hat{u}_k^h \; \hat{z}_k^h\right]^T + \gamma B$
        $h \leftarrow h+1$
    **until** convergence criterion met
    $\hat{\alpha}_k \leftarrow J_{\gamma f}(\hat{\alpha}_k^N, \hat{u}_k^N, \hat{z}_k^N)$
**until** finished
$\hat{X}(k,l) \leftarrow \tilde{Y}(k,l)\hat{\alpha}(k,l)$
```

The convergence criteria of the algorithm is discussed in [6]. In that algorithm, the input γ is chosen to be 0.5 in the experiments. The λ values effect the priority of ℓ_1 and TV norms. In order to achieve better results these values are determined empirically.

Even if λ values are determined empirically, they tend to show a pattern for better estimation with changing input SNR values. It is natural that if input SNR values increase it is more suitable to rely on the observation. Therefore the λ values decrease. With increasing noise presence in the observation, the estimate tends to rely more on constraints. Thus, sparsity and constancy in phase shift becomes increasingly important, which requires increased λ values.

4.3 Experiments and Discussion

In that section, experiments using the proposed method given in Alg.4 are proposed. Through these experiments, the reliability of the method is questioned.

This method is proposed to compensate the effects of the musical noise. As explained before, musical noise is a natural outcome of sparsity based modifications in low input SNR values.

Through the experiments, measured RIRs in a reverberant room are used. Signals are reverberated in some scenarios with the measured impulse response or the RIR measurement is achieved at the same time audio signal is recorded. In both options, it is assumed that RIRs are known.

Tables for output SNR, different λ values for different input SNRs are given in order to demonstrate the effects of the constraints.

In order to justify overall effects, the experiment is repeated for a preset input SNR for the entire signal. Again, λ values are searched in an interval. With changing values of λ the SNR values can differ.With lower SNR values auditory quality can increase. Through experiments it can be observed that with reduced SNR the choice of λ can decrease the effect of musical noise. In each experiment, in order to prove this, the time-frequency coefficients are given for different composition of λ's.

4.3.1 Experiment-1

In this experiment a speech signal is used. A clean audio signal is obtained at first with 44.1[kHz] sampling frequency. The clean signal is synthetically reverbed with the measured room impulse response. Therefore, the SNR can be modified manually. The input SNR is defined on the reverbed signal itself and the noisy observation. The reverberated and noisy observation is used in the proposed method in order to obtain the estimate.

In order to determine the quality of the proposed method, an active frequency band of the reverberated signal is chosen. Complex noise is added to that channel with pre-determined SNR value. The proposed method is applied on that channel in order to obtain the estimate. With different ℓ_1 norm weights (λ_1) and TV norm weights (λ_2) for different input SNR values the experiment is repeated. The results can be checked from Table 4.1. $[\lambda_1, \lambda_2] = [0, 0]$ corresponds to least squares estimation. Output SNR values for LSE are not given as it is determined as the starting point of the iterations. It can be observed that, the output SNR values are lower compared to preset input SNR value. Input SNR ratio is calculated using the noisy reverberated channel and the noiseless reverberated channel, output SNR is calculated using the estimate and the original signal itself. Output SNR comparison also includes the effects of the RIR. Under these conditions, Table 4.1 shows that with the presence of TV norm, SNR increases. This is why $\lambda_2 = 0$ is given in each scenario. This case shows the sparsity solution only. It can also be observed that with the increasing values of λ_1, SNR increases. Thus, with $\lambda_1 = 0$ the results are lower than other cases. This may seem as a contradiction to the importance of the magnitude information. However, as an active frequency band is chosen for the Table 4.1, it is expected to satisfy phase shift to be constant more likely compared to sparsity. Overall, it can be observed that the proposed method yields better results as SNR increases for such scenarios where $\lambda_1 \neq 0$.

Table 4.1: Output SNR Values (Experiment-1)

Input SNR = 12[dB]

	λ_2					
λ_1	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
0		3.18	5.97	7.76	10.26	5.51
10^{-5}	3.89	4.64	6.09	7.78	10.27	5.51
10^{-4}	6.33	6.37	6.65	7.92	10.3	5.51
10^{-3}	8.06	8.07	8.16	8.88	10.55	5.51
10^{-2}	10.81	10.82	10.85	11.11	**11.31**	5.53
10^{-1}	8.72	8.72	8.73	8.81	9.02	5.19

Input SNR = 7[dB]

	λ_2					
λ_1	0	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
0		1.49	2.7	4.42	5.81	4.86
10^{-3}	1.67	2.29	3.2	4.68	5.9	4.87
$10^{-2.5}$	2.79	3.2	3.9	5.13	6.09	4.88
10^{-2}	4.45	4.7	5.15	6.06	6.53	4.92
$10^{-1.5}$	5.85	5.98	6.22	6.72	**6.82**	4.9
10^{-1}	5.69	5.76	5.9	6.18	6.16	4.6

Input SNR = 2[dB]

	λ_2					
λ_1	0	10^{-2}	$10^{-1.5}$	10^{-1}	$10^{-0.5}$	1
0		-1.5	0.39	2.32	2.13	0.62
10^{-2}	-1.3	-0.066	1.34	2.66	2.12	0.61
$10^{-1.5}$	0.68	1.49	2.5	3.16	2.1	0.59
10^{-1}	2.7	3.16	**3.69**	3.56	1.97	0.52
$10^{-0.5}$	3.01	3.24	3.45	3.04	1.58	0.4
1	2.11	2.25	2.38	2.12	1.1	0.22

In order to justify the method, the entire signal is required to be examined. The method is proposed to counter the effects of musical noise. The musical noise components at high frequency components can be encountered with higher λ values. The composition of these values highly depends on the observation.

Effects of the TV and ℓ_1 norms can be observed from Fig.4.1. In Fig.4.1-(c) the estimate is obtained using a moderate level for sparsity and low level for constant phase shift. Thus, the result tend to be sparse. However, this reconstruction still contains unusual activity in high frequency terms, which are the reason of the musical noise. In order to decrease the effects of musical noise the effect of TV norm can be increased. However, increasing both λ values does not increase efficiency. Using a moderate TV norm and low ℓ_1 norm weight gets rid of the musical noise components, however this also yields a small reverb effect in reconstruction. The process can be observed from Fig.4.1-(d).

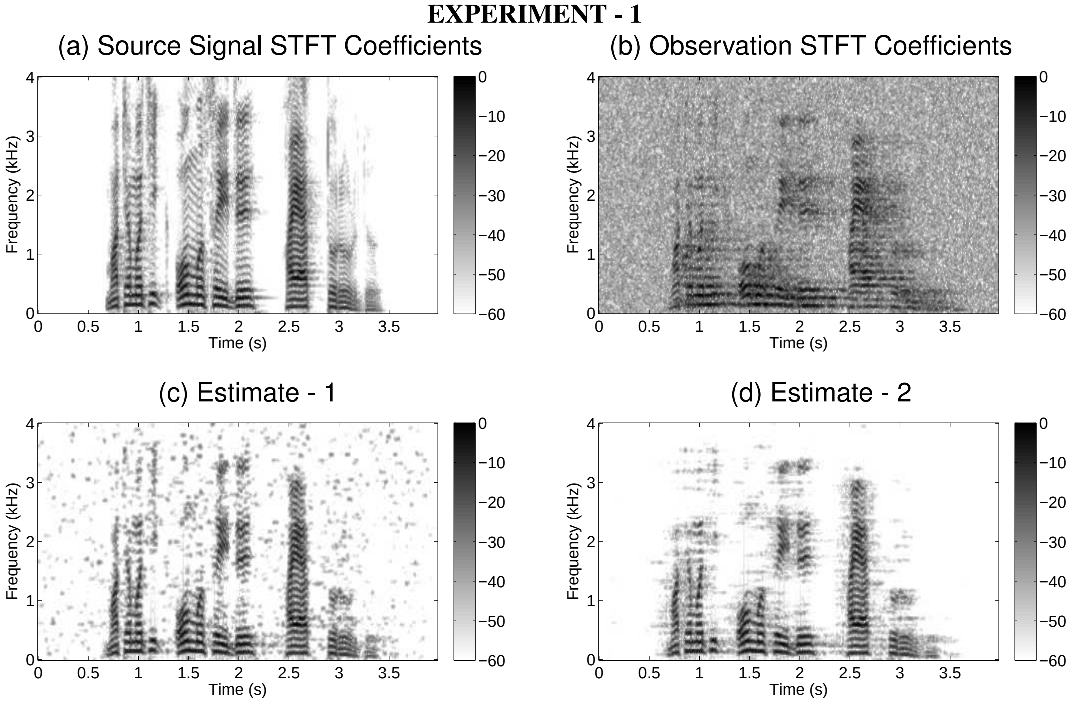

Figure 4.1: (a) Original signal STFT coefficients. (b) Reverbered and noisy observation STFT coefficients (Initial SNR = 12[dB]). (c) Moderate λ_1 and low λ_2 estimate. (d) Low λ_1 and moderate λ_2.

Estimating speech signals with higher ℓ_1 norm priority and lesser TV norm priority for erasing musical noise gives better SNR values in time domain. However, auditory quality is questionable.

4.3.2 Experiment-2

In this experiment, a violin is used to give an example of a musical instrument signal. Same procedure is repeated. The sampling frequency for the signal is chosen as 44.1[kHz]. The original signal is reverberated synthetically. Noise is added satisfying 10[dB] input SNR.

Similar to the first experiment one band SNR values are calculated for time frequency coefficients. It can be seen from the figures that the musical time frequency spectrum has a constant nature. Compared to speech time frequency spectrum it can be assumed that the constant phase shift is reliable. It can be also seen from Table 4.2 that TV norm weight affects SNR gain more compared to sparsity weight. Also it is again justified that employing phase information increases the quality of reconstruction.

It can be observed that, harmonics in coefficients are more distinguishable compared to speech signal given in Fig.4.1. Therefore, the TV norm is expected to have greater impact on estimate. With constant phase shift enforced in mask estimation, the results can be checked from 4.2. Increasing weight of the TV term does not yield dramatic reverb effects which can be compared from Fig.4.1-(d) and Fig.4.2-(d).

Similar to the first experiment, better SNR values in time domain are obtained using moderate λ_1 values and low λ_2 values. However, using a high λ_2 value erases the effects of the musical noise and increases auditory quality.

Table 4.2: Output SNR Values (Experiment-2)

Input SNR = 12[dB]

	λ_2					
λ_1	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
0		7.99	8.4	10.86	13.57	5.48
10^{-5}	8.02	8.08	8.46	10.9	13.57	5.48
10^{-4}	8.52	8.56	8.91	11.19	13.59	5.48
10^{-3}	11.15	11.18	11.44	13.25	13.71	5.47
10^{-2}	14.24	14.25	14.37	**14.88**	13.45	5.34
10^{-1}	9.9	9.9	9.93	10.12	9.96	5.05

Input SNR = 7[dB]

	λ_2					
λ_1	0	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
0		4.64	6.5	9.57	9.28	5.41
10^{-3}	4.79	5.93	7.65	10.14	9.28	5.41
$10^{-2.5}$	6.69	7.76	9.38	10.99	9.27	5.41
10^{-2}	9.3	10.34	11.66	11.82	9.23	5.38
$10^{-1.5}$	10.21	10.91	**11.52**	11.17	8.77	5.16
10^{-1}	8.89	9.25	9.56	9.44	7.74	4.77

Input SNR = 2[dB]

	λ_2					
λ_1	0	10^{-2}	$10^{-1.5}$	10^{-1}	$10^{-0.5}$	1
0		1.49	4.55	4.88	2.33	0.37
10^{-2}	1.95	4.58	6.03	5.05	2.36	0.35
$10^{-1.5}$	5.18	7.06	7.11	5.07	2.42	0.32
10^{-1}	6.25	**7.14**	6.67	4.63	2.49	0.25
$10^{-0.5}$	4.97	5.88	5.71	4.2	2.29	0.098
1	3.62	4.1	4.23	3.05	1.58	0.005

EXPERIMENT - 2

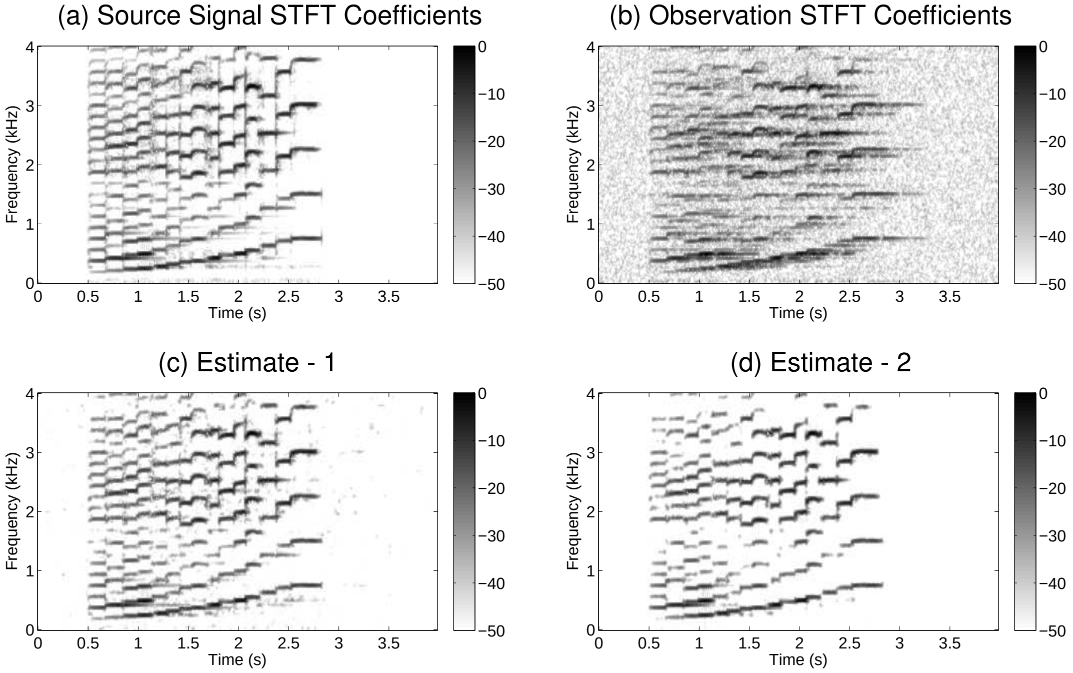

Figure 4.2: (a) Original signal STFT coefficients. (b) Reverbered and noisy observation STFT coefficients (Initial SNR = 10[dB]). (c) Moderate λ_1 and low λ_2 estimate. (d) Low λ_1 and moderate λ_2.

5. MIXED NORM REGULARIZATION

Dereverberation using single observation channel is demonstrated with sparsity constraint in chapter 3 and additionally, employing phase information in chapter 4. In applications microphone arrays are used for audio recordings which yield multiple observations. It is expected to increase the estimation quality. Because with more observations sharing the same information, it is natural to assume that the original information can be obtained explicitly.

In most cases, the array geometry is tried to be exploited using the phase shift in observations. In denoising cases, mostly delay and sum beamformers are used in order to obtain a clean signal [18, 19]. In derevereberation case, however, room impulse responses vary dramatically due to location. This limits the usage of the geometry, as the reverb operators differ with source location. With the unstable nature of RIR, multiple observations can be taken into consideration together. It is observed that observations recorded close enough share a closely related information about the source. This information can be extracted from these observations. This also poses the idea that a common filter can be defined for the recording environment. Obtaining RIRs for each observation location is problematic. Instead, preliminary experiments can be performed in order to obtain a common filter that forms the RIRs. Using that common filter the RIR can be assumed to be known. Also this common filter can be used in order to shorten the effects of reverberation.

With known impulse responses for the multichannel case, a convex optimization problem can be posed. The convex problems are often address to multichannel convex penalty functions using principle components of the sources. [20–24]. It is assumed that, as the source is active in observations, the observations share a common information. Different from these methods, the penalty function explained and used in this chapter is mixed norm [25]. Mixed norm penalty function uses time frequency coefficients of each observation, where principle component investigation requires an extraction. Thus, without a special decomposition required, mixed norm basically checks the similarities in observations. This penalty function, for a fixed time and a fixed frequency, checks the similarities in a microphone vector which consists time-frequency coefficients.

The microphones in an array are located close to each other (1-2[cm] away from each other). Therefore, wave front of the acoustic signal does not reach microphones with dramatic difference. Each of the observations are in similar in pattern. Therefore the time shift between observations is expected to be considerably little. If the STFT of these signals are taken, in a specified frequency band, corresponding time coefficients show similar activity. Thus, a harmonic is expected to be found in the other observations with a little time shift. This information is used in order to separate noise with the source activity in the spectrum. It is known that the source signal appears in each observation with different RIR effects. However, source observation is common in each observation which results in similarities. For dereverberation case, effects of the reverberation can be shortened assuming that RIR filters are different. Also noise terms are not expected to be active in all microphones at the exact time frequency coefficient. With mixed norm regularization, it is checked if a time frequency coefficient is active in all observation spectra. Using the vectors for fixed a time and fixed a frequency, it is checked if for this time-frequency bin enough of the observations are active. It is also considered that with the effects of reverberation and noise presence one time-frequency does not give enough information. Instead of checking sparsity of the spectrum, sparsity of the harmonics in the spectrum can be investigated. In order to investigate the presence of a harmonic in a frequency band, neighboring coefficients are also taken into consideration by using a block form of coefficients. With forming a block of coefficients in a frequency band, harmonic structure is tried to be recognized. With grouping coefficients, it is desired to limit of harmonics in a frequency band. The number of harmonics are tried to limited by checking the ℓ_2 norm of the blocks. This addresses to a new penalty function, that groups up the coefficients, in order to check if the activity in a time frequency coefficient belongs to a harmonic. Audio denoising is achieved using block thresholding in [26]. In this chapter, multichannel data is grouped up.

In a harmonic of a specified frequency band, time sequential coefficients are active. In order to check whether the coefficient belongs to the harmonic structure, the values of sequential coefficients together are taken into consideration. Different from mixed norm regularization, this process groups up the coefficients in order to form a block. Block mixed norm regularization calculates mixed norm over coefficient blocks. In this regularization for a specified frequency, time axis is divided into non-overlapping groups of the same size. Each of these portions form a matrix of time-microphone coefficients. In here, it can be observed that, with the presence of a harmonic, the matrix coefficients tend to be active. The harmonic is also expected to be observed in each observation. Thus, it can be said that the matrix is expected to have a high ℓ_2 norm. Using this information the time interval can be suppressed if its ℓ_2 norm is below a threshold. Thus, with limiting the number of these blocks, effects of reverberation is shortened. As the reverberation effect in each microphone is different and the RIRs are modelled as decreasing sequences, active coefficients caused by reverberation tend to have low ℓ_2 norms.

In the next section, the generalized multichannel problem is defined in STFT domain. This definition is mandatory because this chapter explains both denoising and dereverberation using the same penalty function. The genaralized definition can be compared with the definition proposed in Eq.(2.4). This chapter has the generalized problem definition entirely on STFT coefficients. In this chapter mixed norm and block mixed norm are defined. The algorithm for solving the convex optimization processes for denoising and dereverberation are proposed. Also the quality for both scenarios are questioned with the experiments.

5.1 Proposed Method

In this chapter multiple microphone case is taken into consideration. In multichannel signal processing, the observations are assumed to be effected by different filters and the noise. The general multichannel signal observation in time domain can be modeled as,

$$y_m = H_m x + n_m \tag{5.1}$$

y_m is the observation signal at m^{th} microphone. H_m's are the convolution operator with the filters for each channel. If all H_m's are assumed to be I the model becomes a model for denoising problem, otherwise we have to solve the multichannel deconvolution problem. n_m's are the additive Gaussian noise for each channel.

It is assumed that the room impulse response can be expressed in STFT domain, which is explained in Chapter 3. With this assumption the problem can be defined on time frequency coefficients. The problem can be rewritten,

$$Y_m = \mathscr{H}_m X + U_m \tag{5.2}$$

Here in this definition, $\mathscr{H}_m$ denotes reverberation operator for corresponding observation. As the proposed filters represent the effects of RIR, the estimation can be formed for each frequency band. Thus, the problem can be solved separately.

In order to revert the effects of reverberation and remove noise, a convex optimization problem can be used. In order to estimate the signal of interest, a minimization can be performed over penalty functions. The minimization process of the penalty function in general can be given as,

$$\min_X \left\{ g_m(X) = \|Y_m - \mathscr{H}_m X\|_2^2 + \sum_i \lambda_i p_i(X) \right\} \tag{5.3}$$

In this notation $p_i(\cdot)$ defines the penalty functions with weights λ_is. For example ℓ_1 norm enforces sparsity by penalizing magnitude distance from 0.

The solution to this problem is troublesome, because the room impulse responses differ dramatically for each microphone with a constant source signal. Instead of solving this problem in order to explicitly obtain the source, the reverberation effects can be shortened. This equivalent problem can be formed with a relatively short impulse response definition. The problem has a common filter that maps different sources to corresponding observations. In order to pose such a problem, relatively short impulse response concept is proposed. The shortened impulse response divides the RIRs into a common part and the independent parts. The common part is assumed to be same for each RIR where independent parts are the residuals. The residuals can be inserted into sources in order to leave common filter being the same reverberation filter for each observation. Therefore, the problem becomes estimating different sources (source convolved with residual RIRs) sharing common information with the same reverberation filter (common RIR). This modification allows the usage of mixed norm in the problem. Within this multichannel estimation, the aim is to preserve harmonic structure.

The next section explains the relatively short RIR concept.

5.1.1 Relatively short impulse response

It is explained that, in an environment the RIR is hard to determine explicitly for each observation location. Therefore with preliminary experiments a shorter but a common filter that represents the RIR can be found. With that model assumption common RIR of the environment can be considered as the only filter. Instead of using one source signal and different filters, it is aimed to have different sources with the same reverberation filter. Using the shortened RIR definition, the reverberation filters can be shortened, which also allows to use this common filter to define moderately reverberated observations. These observations can be treated as different source signals, which share the common source signal. Shortened impulse response concept can be replaced with relative transfer function, where the common part of the RIR is assumed to be one of the RIRs. The residuals are set to the relative transfer functions. However, for this application shortened impulse response is used. For further reading about relative transfer function or shortened impulse response concepts [27–30]can be checked.

RIR differs dramatically according to the position of the microphone. Room impulse response is expressed as two filters. The common filter, is independent of location and the same for all observation points. The second filter is position dependent which is different for each microphone location. Let $*$ denote the convolution operator, the relation is defined as,

$$y_m = h_m * x \approx gc * gi_m * x = gc * z_m \tag{5.4}$$

In the equation Eq.(5.4), gc denotes the common filter of the room impulse response independent of the location, gi_ms denote the independent filter of the room impulse response differing for each observation point. Modified impulse response can be inserted into the original problem. Using modified impulse responses, reverberation effects can be shortened. Observations with shortened RIR is denoted as z_m in Eq.(5.4). These modified signals are assumed to have different sources for each observation. Therefore, the problem given in Eq.(5.5) turns into,

$$\min_{Z} \left\{ g(Z) = \|Y - \mathscr{G}Z\|_2^2 + \sum_i \lambda_i p_i(Z) \right\} \tag{5.5}$$

Here $\mathscr{G}$ is the common filter for each channel. Where $Y = [Y_1, Y_2, \cdots Y_M]^T$ and $Z = [Z_1, Z_2, \cdots Z_M]^T$ are the STFT coefficients of y_m's and z_m's respectively. Here $\mathscr{G}$ is obtained as the LSE of a linear system. Using the shortened impule response relation the length of the common and residual parts are determined previously.

With the help of relatively short room impulse responses, single channel estimation turns into multichannel estimation. With the definition of $\mathscr{G}$, instead of estimating X only, the aim is to estimate Z's, where Z's denote the moderately reverberated versions of X. Therefore, this gives the independence of solving the problem with separating for each observation.

Different from Eq.(5.1) , here the overall penalty function is multichannel. Therefore, the penalty functions $p(\cdot)$ have the freedom of penalizing multichannel data.

In the next section Mixed Norm is discussed as the multichannel penalty function.

5.2 Mixed Norm

Assume that the multichannel time-frequency spectrum can be considered as a three dimensional data. Window employed in STFT is 60[ms] long, which is a common length for audio signals. Also, microphone array is uniformly distributed and the distance between two corresponding microphone is relatively low. Thus, this can be assumed that the 60[ms] differences in time domain does not make marginal differences in STFT domain. Also it is known that in order to have this long difference the microphones are required to be afar. In the given geometry microphones are closely located. Under these conditions there can not be marginal differences between observation STFT spectra. Therefore, as a natural result, if a time frequency coefficient is active in a microphone coefficient it is expected to be active in other microphones as well. Consequently, if there is an activity caused by the source in a specific time frequency bin of the coefficients it is expected to have activity in other microphones as well. Therefore this relation can be penalized.

Vectors of interest are of length M for each time-frequency coefficient, where M denotes the microphone number. It can be observed that the mixed norm uses all provided observations If the vector has a high ℓ_2 norm it is assumed that there exists an activity as it exists in all observations. The mixed norm enforces sparsity on time-frequency plane employing the ℓ_2 norm of these vectors. It is desired to have a limited number of active vectors in time-frequency plane. Mixed norm penalty function is defined as,

$$p(Z) = \|Z\|_m = \left| \sum_{k,l} \left(\sum_q |Z_q(k,l)|^2| \right)^{\frac{1}{2}} \right| \tag{5.6}$$

Here, depending on the ℓ_2 norm of microphone vectors for each time frequency coefficient, the sparsity constraint is enforced. To explain, this constraint checks the total activity in each time frequency bin, using all observations. The final penalty function can be defined as,

$$\min_Z \|Y - \mathscr{G}Z\|_2^2 + \lambda \|Z\|_m \tag{5.7}$$

In addition, it is mandatory to mention that $Y = [Y_1 \, Y_2 \cdots Y_M]^T$ here is a vector which consist all observations. $Z = [Z_1 \, Z_2 \cdots Z_M]^T$ is the source convolved with relative transfer function. $\mathscr{G}$ is the convolution operator with the common part of the shortened impulse response.Relatively short transfer function is used for denoising in [31]. Differenty from the given scenario, multichannel data is penalized with a penalty function.

This problem can be solved using DR Algorithm. Assuming that the penalty function can be divided into two parts. The penalty function can be divided as,

$$\begin{aligned} f(Z) &= \|Y - \mathscr{G}Z\|_2^2 \\ g(Z) &= \lambda \|Z\|_m \end{aligned} \tag{5.8}$$

In order to form DR iterations, the proximity operators of these functions are required.

Proximal calculation

Proximal of the quadratic term can be calculated as,

$$J_{\gamma f}(Z) = \arg\min_Q \frac{1}{2\gamma} \|Z - Q\|_2^2 + \frac{1}{2} \|Y - \mathscr{G}Q\|_2^2 \tag{5.9}$$

As the problem is convex, it can be said that 0 is an element of the gradient of this problem. Assume that $\hat{Q}$ minimizes the problem, solution can be obtained. By taking the gradient and setting it to 0 as,

$$\begin{aligned} 0 &\in \left(\hat{Q} - Z\right) + \gamma \mathscr{G}^* \left(\mathscr{G}\hat{Q} - Y\right) \\ J_{\gamma f}(Z) &= \left(\gamma \mathscr{G}^* \mathscr{G} + I\right)^{-1} \left(\gamma \mathscr{H}^* Y + Z\right) \end{aligned} \tag{5.10}$$

Proximal of the mixed norm can be found using,

$$J_{\gamma g}(Z) = \arg\min_{Q} \frac{1}{2}||Z - Q||_2^2 + \lambda\,||Q||_m \tag{5.11}$$

It can be observed from Eq.(5.6), that the mixed norm can be considered as a ℓ_2 norm calculation for each microphone vector for specified time frequency coefficient. Therefore, the proximity function of mixed norm can be calculated as ℓ_2 norm proximity functions for each time frequency coefficient. Let $Z_{l,k}$ denote the $1 \times M$ microphone vector at the l^{th} time and k^{th} frequency bin and let $(J_g(Z))_{l,k}$ denote the value of the proximal at the l^{th} time and k^{th} frequency bin. The proximal function is,

$$\left(J_{\gamma g}(Z)\right)_{l,k} = \frac{\text{soft}\left(||Z_{l,k}||_2, \lambda\right)}{||Z_{k,l}||_2} Z_{k,l} \tag{5.12}$$

This statement can be checked from Eq.(2.17), where the definition is obtained using positive values of the scaling factor. If this factor is generalized the result is the soft thresholding operator. The proximals can be inserted into DR Algorithm. The solution is,

Algorithm 5 DR-Solution for Eq.5.7

```
$\gamma \in (0,1)$, $\lambda \in \mathbb{R}^+$
$\hat{Z}(l,m) = 0, A(l,m) = 0, B(l,m) = 0\ \forall l,m$
repeat
    $A \leftarrow 2J_{\gamma f}(\hat{Z}) - \hat{Z}$                Eq.(5.10)
    $B \leftarrow J_{\gamma g}(A)$                              Eq.(5.12)
    $\hat{Z} \leftarrow (1-\gamma)\hat{Z} + \gamma(2B - A)$
until convergence criterion met
$\hat{Z} \leftarrow J_{\gamma f}(\hat{Z})$
```

This solution may be improved by modifying the penalty function as explained before. In order to increase the efficiency. Block mixed norm regularization is defined in the next subsection.

5.2.1 Block mixed norm regularization

In this section Block mixed norm is taken into consideration which is visualized in Fig.5.1.

Mixed norm treats the time frequency coefficients over microphones as a vector. Thus, this only takes one coefficient vector into consideration. In STFT domain the coefficients tend to be closely valued in magnitude especially in harmonics. Instead

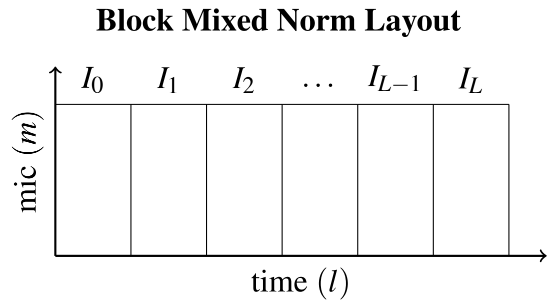

Figure 5.1: For a specified frequency band, l denotes the time axis, m denotes the microphone axis. Here each I_i has the same size.

of comparing all time frequency bins one by one, it is more suitable to group up time consecutive coefficients in a specified frequency band and check for harmonic existence in this interval. This idea is also employed in single observation denoising application in [26]. Therefore, the nature of the harmonics are aimed to be preserved even when there is a slight time shift between microphones. Instead of $1 \times M$ vectors the coefficients of interest here form a matrix of size $K \times M$. Consider that I_i denotes the i^{th} interval of length K in the frequency band. Therefore, the length of the signal is required to be integer multiple of K for non-overlapping blocks. With the assumption X_k denotes k^{th} frequency band of the time frequency spectrum X, block mixed norm on a specified k^{th} frequency band can be denoted as,

$$\|X_k\|_{K,m} = \left| \sum_i \left(\sum_{q,l \in I_i} \|X(k,l,q)\|^2 \right)^{\frac{1}{2}} \right| \quad \text{where } [I_i]_{K \times M} \tag{5.13}$$

With the block mixed norm definition, the problem defined in Eq.(5.7) can be modified. In order to apply the block mixed norm procedure, a specific frequency band should be chosen. Let Z_k denote a specified frequency band of the STFT coefficients. Then the problem can be formed as,

$$\hat{Z}_k = \arg\min_{Z} \frac{1}{2} \|Y - \mathcal{G} Z_k\|_2^2 + \lambda \|Z_k\|_{K,m} \tag{5.14}$$

In this problem definition one frequency band at a time is taken into consideration. Therefore, the overall solution can be obtained after solving this problem for entire frequency vectors.

The penalty function given in Eq.(5.14) can be solved using DR Algorithm. The algorithm requires the proximal values of the functions. Proximal of the quadratic term is analogous to the previous problem and can be checked from Eq.(5.10).

Here, the mixed norm is defined on the small intervals on the frequency band. It is assumed that these intervals, which are $K \times M$ matrices, are functions of microphone and time. It can be observed that the proximity operator of the blocked mixed norm can be done for each block separately. In this problem the matrix is penalized with ℓ_2 norm. Here, different from mixed norm regularization, a matrix is taken into consideration. Let $\tilde{O}$ denote the matrix of size $K \times M$ visualized in Fig.5.1 and $\tilde{O}_i$ the i^{th} matrix of interest. Thus, the calculation is again can be done analogous to the vector scenario. The proximal of the function for the matrix $\tilde{O}_i$ can be calculated as,

$$J_{\gamma g}(\tilde{O}_i) = \frac{\text{soft}\left(\|\tilde{O}_i\|_2, \gamma\right)}{\|\tilde{O}_i\|_2} \tilde{O}_i \tag{5.15}$$

This proximity calculation is required to be done for each interval. This form can be generalized for the entire frequency band. Thus, each proximal calculated for the matrix, is the element of the specified frequency band. The proximal of the frequency band can be expressed as,

$$J_{\gamma g}(Z_k) = \left[J_{\gamma g}(I_1), J_{\gamma g}(I_2), \cdots, J_{\gamma g}(I_i)\right]^T \tag{5.16}$$

With the knowledge of proximity functions, the penalty function can be minimized using DR algorithm. The Algorithm for minimizing the problem in Eq.(5.14) is analogous to the problem in Eq.(5.7). Overall, the complete algorithm can be proposed with the help of Alg.5 as,

Algorithm 6 Mixed Norm Penalized Dereverberation

```
$\gamma \in (0,1)$, $\lambda \in \mathbb{R}^+$, $K \in \mathbb{Z}^+$
$\hat{Z}(l,k,m) = 0$, $A(l,k,m) = 0$, $B(l,k,m) = 0 \; \forall l,k,m$
repeat $\forall k$
    repeat
        $A \leftarrow 2J_{\gamma f}(\hat{Z}_k) - \hat{Z}_k$                Eq.(5.10)
        $B \leftarrow J_{\gamma g}(A)$                                   Eq.(5.15),(5.16)
        $\hat{Z}_k \leftarrow (1-\gamma)\hat{Z} + \gamma(2B - A)$
    until convergence criterion met
    $\hat{Z}_k \leftarrow J_{\gamma f}(\hat{Z}_k)$
until finished
```

The algorithm steps for blocked mixed norm regularization can be performed easily. However, the matrix inversion, considering the inversion of RIR effects which denoted in Eq.(5.10), is time consuming. In order to decrease required computational time, the trick is to make us of the Taylor Series expansion,

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \tag{5.17}$$

In that sense the matrix inversion, if $I-(-1)\alpha\mathscr{G}^T\mathscr{G}$ is assumed, can be calculated as infinite multiplication. Thus the inversion can be stated,

$$(I+\alpha\mathscr{G}^*\mathscr{G})^{-1}=\sum_{n=0}^{\infty}(-1)^n\,\alpha^n\,(\mathscr{G}^*\mathscr{G})^n \tag{5.18}$$

In this inversion, the $\mathscr{G}$ denotes the convolution matrix with the common room impulse response in STFT domain. A convolution matrix can be denoted with a toeplitz matrix, if the signal to be convolved is extended to the observation length [32]. Thus, the toeplitz matrix multiplied with the transpose form a symmetric matrix with non zero diagonal. Therefore, the singular value decomposition will yield FT coefficients as singular values and the FT as the singular vectors.

Then, taking the n^{th} power corresponds to taking FT, taking the power of the roots then taking the inverse transform. The inversion stated in Eq.(5.18) becomes,

$$(I+\alpha\mathscr{H}^*\mathscr{H})^{-1}=C^{-1}\left(\sum_{n=0}^{\infty}(-1)^n\alpha^n\Sigma^n\right)C \tag{5.19}$$

Where C and C^{-1} denotes the FT and its inverse. Therefore the multiplication with the inverse matrix can be represented as a circular convolution. The convolution can be performed as multiplication in time domain. Thus the computational time decreases. In Alg.6 'A' can be calculated entirely in Fourier domain. However, since mixed norm is applied on the time axis for specified frequency bin, transform and its inverse is required. This is because FT is applied on each frequency band.

It can be observed that the algorithm proposed in Alg.6 can be used for both dereverberation and denoising. Assuming that H_m's defined in Eq.(5.1) are identity matrices, the problem directly denotes the denoising. Analogously, shortened impulse response becomes equal to the identity and the observations with different noise are used in the algorithm. Denoising algorithm can be derived from dereverberation algorithm just by considering all the impulse responses are identity. Simplifications are mandatory in this process, because, with all filters are considered to be identity, the convolution operators are not required. Also the matrix inversion becomes division with a real number.

Mixed norm regularization is used for both denoising and dereverberation purposes. In the next section experiments are demonstrated and the advantage of blocked mixed norm regularization is tried to be explained.

5.3 Experiments

In this section the denoising and dereverberation experiments are proposed. As explained in the previous sections, denoising application uses the same algorithm as assuming the room impulse response as $\delta(n)$. With suitable simplifications $\mathscr{H}$, the modified impulse response is equal to the identity matrix. In the next subsections experiments for denoising application and dereverberation application is taken into consideration separately.

5.3.1 Denoising experiment

In order to justify the quality of the method proposed in Alg.6, a multichannel data is created using different noises for each observation. A speech signal is taken with 44.1[kHz] sampling frequency and Gaussian noise is added synthetically. The aim in this application is to estimate the original signal in each channel with reduced noise.

Observations are formed synthetically with different input SNR values for each experiment. Regularization takes two input values for this scenario: the filter length and the thresholding parameter. In order to justify that, using block mixed norm regularization yields better results compared to conventional mixed norm regularization, experiments on an active frequency channel are performed. Experiments are performed using 4 and 8 microphone cases.

Table 5.1: Denoising Output Gains.

		K						
mic.	**SNR**	1	3	5	7	9	11	13
4	0	4.49	4.96	**5.15**	4.93	4.78	5.06	4.91
4	5	2.74	2.97	**3.36**	3.22	2.98	3.12	3.12
4	10	2.077	2.218	**2.337**	2.329	2.207	2.092	2.134
4	15	1.476	1.614	1.554	**1.623**	1.575	1.529	1.429
8	0	4.63	4.86	4.91	**5.11**	4.88	4.64	4.38
8	5	3.23	3.36	3.2	**3.41**	3.31	3.16	2.94
8	10	2.253	**2.338**	2.149	2.283	2.301	2.226	2.09
8	15	1.476	**1.524**	1.488	1.396	1.498	1.472	1.418

For corresponding input SNR values and for different number of microphones, SNR gains are calculated for different values of filter length K. The SNR gains are the best values for different λ values. The results can be checked from Table 5.1. $K = 1$ defines the conventional mixed norm regularization. It can be said that using a block mixed norm penalty function increases the efficiency of regularization. Also, as expected with increasing the input SNR values the gain decreases.

In order to visualize the effects of the algorithm on the time frequency coefficients, the speech signal is used. With input SNR 5[dB] and 4 observations, the proposed method is used in order to achieve the best gain. The parameters λ and K are chosen empirically after some experiments. The results are visualized in Fig.5.2.

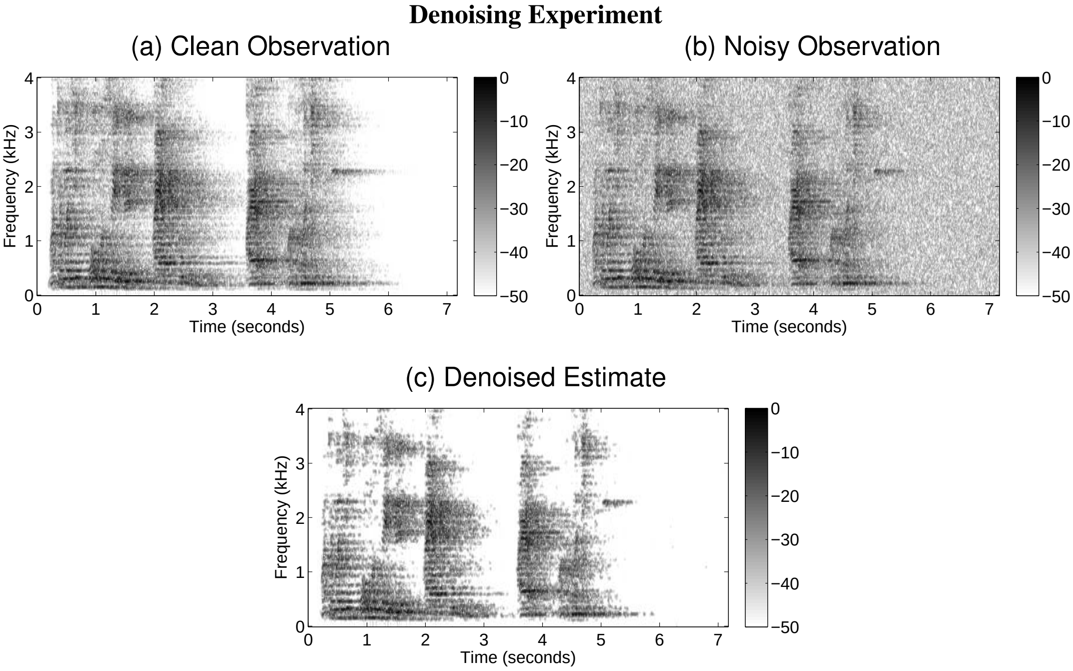

Figure 5.2: (a) Original observation at mic. number 4. (b) Noisy observation. (c) Estimate using proposed method. There are 3 more reconstruction examples because the method is applied on 4 microphone case.

The estimate is obtained 7.12[dB] gain from the noisy observation. In this experiment, the SNR value is calculated in time domain comparing the estimate and the clean signal. Different from this experiment, Table 5.1 is formed calculating SNR using this frequency vector which explains the difference between best obtained SNR values of one channel case and this experiment.

5.3.2 Dereverberation experiment

In order to perform dereverberation, measured impulse responses in the lab environment are used. The microphones for room impulse response measurement are in an array layout with 3[cm] distance between each. The room impulse responses are of length 1.5[s]. As explained before in order to form closely related observations, it is required to calculate the shortened RIR. The common part of the RIR is factorized with length 1[s].

In order to justify using block mixed norm regularization over conventional mixed norm regularization a series of experiments are performed. In the experiments an active frequency band of time-frequency transform is used. The band is chosen from reverberated signals of each microphone and complex noise is added with previously set SNR values. With the knowledge of common impulse response, the moderate reverbed estimates are obtained. SNR values are obtained comparing the estimate band and the original moderate reverbed band.

Table 5.2: Dereverberation Output SNR

		K						
mic.	**SNR**	1	3	5	7	9	11	13
4	0	4.57	5	5.14	**5.14**	4.95	5.08	5.08
4	5	7.84	8.27	8.15	8.26	8.35	**8.36**	8.18
4	10	11.4	11.83	**11.98**	11.81	11.9	11.96	11.88
4	15	15	15.51	**15.87**	15.69	15.74	15.77	15.77
8	0	4.89	5.08	5.14	4.86	4.96	**5.19**	5.04
8	5	7.84	8.04	**8.33**	8.18	8.05	8.19	8.18
8	10	11.63	11.81	**11.99**	11.91	11.83	11.76	11.77
8	15	15.52	15.75	15.91	**15.93**	15.83	15.8	15.65

Different from denoising experiments, the SNR gain is irrelevant in this situation since there exists a linear operator between the observation and the original signal. The results can be checked from Table 5.2. It can be observed, by making the same comparison in previous section, that block mixed norm regularization has better performance compared to conventional method.

In order to visualize the effects of the algorithm on the time frequency spectrum, the speech signal is used again. In order to highlight the effects of dereverberation, input SNR is chosen as 15[dB] which is calculated using reverberated observation and noisy and reverberated observation. The reconstruction is obtained with SNR 20[dB]. Output SNR is calculated using moderate reverberated observation and the estimate. Removing the effects of the common part of the room impulse response is visualized in Fig.5.3. It can be observed that, using the proposed method shortens the effects of the impulse response by removing the effects of the common part of the impulse response. It can also be stated that, the common part of the RIR is relatively long compared to the independent parts. Therefore, the auditory quality increases.

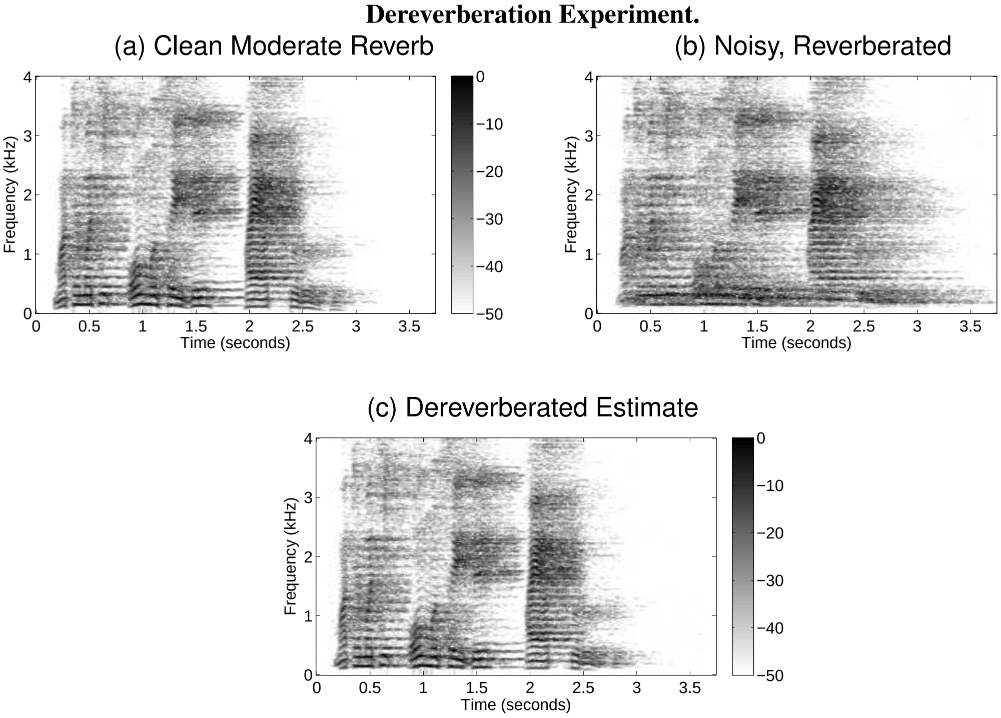

Figure 5.3: (a) Original observation at mic. number 4. (b) Noisy observation. (c) Estimate using proposed method. There are 3 more reconstruction examples because the method is applied on 4 microphone case.

6. CONCLUSIONS AND RECOMMENDATIONS

In this thesis dereverberation problem is taken into consideration where RIRs are known. The non-blind deconvolution problem is solved with the help convex optimization problems using different types of penalty functions.

In chapter 3 sparsity based derevereberation is taken into consideration. Sparsity of time frequency coefficients is enforced using ℓ_1 norm in STFT domain. However, time domain expression is also required to work with RIR and its inverse. This induces domain changes in solution. As the solution is obtained iteratively, computational load increases dramatically with changing domain twice in each iteration. As it is explained in the chapter, in order to avoid domain changes in iterations, RIR is represented as a convolutive operator in STFT frequency bands. This representation is justified with the experiments. With the representation, convex minimization problem is defined in STFT domain. Compared to conventional sparsity based methods, proposed method achieved the similar results with reduced computational time. Considering these benefits, RIR representation is used through the thesis. However, it is also observed that assuming sparsity of the coefficients alone yields musical noise. Increasing the weight of the sparsity constraint in the minimization is not a solution, even it removes the musical noise. Increasing the threshold, also damages the harmonics and decreases quality.

Solution to the musical noise problem is defined in chapter 4. It is observed that only modifying the magnitudes of time frequency coefficients does not yield a perfect solution. In order to increase the efficiency phase information is exploited. Compared to magnitude information, phase information in STFT coefficients is not robust to noise and reverberation effects. Also phase information is reliable only on harmonics which is not that fragile. However, there is a constant phase shift between coefficients in harmonics. This information is used to define a mask between the signal frequency bands and phase shifted frequency bands. Therefore, an optimal mask can increase the estimate quality when applied on the LSE. It is also stated in the chapter that, the masks for each frequency band are required to be piece wise constant. The constant nature of the mask is expected to preserve the harmonic structure, because harmonics are formed by consecutively active coefficients. Musical noise terms however are assumed to be erased, because the coefficients are not correlated. Thus the experiments section in the chapter shows that an optimal composition of constraint weights preserves harmonic structure while erasing the musical noise components. The weight coefficients are chosen to be same for each frequency band mask calculation. However, it can be observed that in active frequency bands, as expected, the vectors diverge from being sparse. Thus, the phase constancy is more valid in these vectors. In order to get better results, weight composition can be modified with taking band activity in consideration. Active bands satisfy being constant more than being sparse. Inactive bands on the other hand are strictly sparse. With a weight factor depending on the band activity the quality is expected to increase. This is aimed to be solved in future.

In chapter 3 and chapter 4 the dereverberation problem is solved with a single observation. In chapter 5, derevereberation problem with multiple microphone case is taken into consideration. In that chapter multichannel derevereberation problem is defined in STFT coefficients using relatively short RIR definition. In multichannel model, observations are formed from the same source with different RIRs and noises. However, with the given geometry of the microphone array, it can be observed that the observations are not disperse. It is observed that, in STFT domain harmonics do not shift dramatically in time between observations. Thus this property addresses that if a time frequency coefficient is active in all observation spectra, it is supposedly a harmonic component. In order to use that property, instead of estimating the source explicitly, the effects of reverberation is aimed to be shortened. For that purpose, it is considered that RIR is the same for all observations where the sources are different. With the definition of relatively short RIR definition, the reverberation operator is divided into two parts. Using the common part only, the problem is modified as a multiple source estimation problem. In that scenario the signals of interest are the moderately reverberated signals (independent RIR convolved with the source). each time frequency coefficient forms a vector with the same coefficient in other observations. As it is assumed that the spectrum is sparse, the number of these vectors should be limited. In order to satisfy this property mixed norm is defined. In order to include the nature of the harmonics blocked mixed norm regularization is defined. Both definitions are supported and questioned with experiments. In the experiments section denoising and dereverberation applications are proposed. For future step, mixed norm regularization is aimed to be used in a multichannel source separation problem.

REFERENCES

[1] **Naylor, P.A. and Gaubitch, N.D.** (2010). *Speech dereverberation*, Springer Science & Business Media.

[2] **Huang, Y.A., Benesty, J. and Chen, J.** (2005). A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment, *Speech and Audio Processing, IEEE Transactions on*, *13*(5), 882–895.

[3] **Furuya, K.I. and Kataoka, A.** (2007). Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction, *Audio, Speech, and Language Processing, IEEE Transactions on*, *15*(5), 1579–1591.

[4] **Yoshioka, T., Nakatani, T., Miyoshi, M. and Okuno, H.G.** (2011). Blind separation and dereverberation of speech mixtures by joint optimization, *Audio, Speech, and Language Processing, IEEE Transactions on*, *19*(1), 69–84.

[5] **Cohen, L.** (1989). Time-frequency distributions-a review, *Proceedings of the IEEE*, *77*(7), 941–981.

[6] **Eckstein, J. and Bertsekas, D.P.** (1992). On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators, *Mathematical Programming*, *55*(1-3), 293–318.

[7] **Petersen, K.B., Pedersen, M.S.** *et al.* (2008). The matrix cookbook, *Technical University of Denmark*, *7*, 15.

[8] **Kowalski, M., Vincent, E. and Gribonval, R.** (2010). Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation, *Audio, Speech, and Language Processing, IEEE Transactions on*, *18*(7), 1818–1829.

[9] **Reilly, J.P., Wilbur, M., Seibert, M. and Ahmadvand, N.** (2002). The complex subband decomposition and its application to the decimation of large adaptive filtering problems, *Signal Processing, IEEE Transactions on*, *50*(11), 2730–2743.

[10] **Combettes, P.L. and Wajs, V.R.** (2005). Signal recovery by proximal forward-backward splitting, *Multiscale Modeling & Simulation*, *4*(4), 1168–1200.

[11] **McAulay, R.J. and Quatieri, T.F.** (1986). Speech analysis/synthesis based on a sinusoidal representation, *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *34*(4), 744–754.

[12] **Siedenburg, K. and Dörfler, M.** (2011). Structured sparsity for audio signals, *Proceeding of 14th conference on digital audio effects (DAFx).*

[13] **Smith, J.O. and Serra, X.** (1987). *PARSHL: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation*, CCRMA, Department of Music, Stanford University.

[14] **Bayram, I.** (2014). Employing phase information for audio denoising, *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE, pp.2893–2897.

[15] **Breiman, L.** (1995). Better subset regression using the nonnegative garrote, *Technometrics*, *37*(4), 373–384.

[16] **Combettes, P.L. and Pesquet, J.C.**, (2011). Proximal splitting methods in signal processing, Fixed-point algorithms for inverse problems in science and engineering, Springer, pp.185–212.

[17] **Lions, P.L. and Mercier, B.** (1979). Splitting algorithms for the sum of two nonlinear operators, *SIAM Journal on Numerical Analysis*, *16*(6), 964–979.

[18] **Talmon, R.**, **Cohen, I. and Gannot, S.** (2009). Multichannel speech enhancement using convolutive transfer function approximation in reverberant environments, *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, pp.3885–3888.

[19] **Schwartz, B.**, **Gannot, S. and Habets, E.A.** (2013). Multi-microphone speech dereverberation using expectation-maximization and Kalman smoothing, *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, IEEE, pp.1–5.

[20] **Shi, Z.**, **Han, J. and Zheng, T.** (2011). A novel framework based on trace norm minimization for audio event detection, *Neural Information Processing*, Springer, pp.646–654.

[21] **Shi, Z.**, **Han, J. and Zheng, T.** (2013). Audio classification with low-rank matrix representation features, *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 15.

[22] **Gu, S.**, **Zhang, L.**, **Zuo, W. and Feng, X.** (2014). Weighted nuclear norm minimization with application to image denoising, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2862–2869.

[23] **Ahmed, A.**, **Recht, B. and Romberg, J.** (2014). Blind deconvolution using convex programming, *Information Theory, IEEE Transactions on*, *60*(3), 1711–1732.

[24] **Cai, J.F.**, **Candès, E.J. and Shen, Z.** (2010). A singular value thresholding algorithm for matrix completion, *SIAM Journal on Optimization*, *20*(4), 1956–1982.

[25] **Wang, J.**, **Liu, J. and Ye, J.** (2013). Efficient mixed-norm regularization: algorithms and safe screening methods, *arXiv preprint arXiv:1307.4156*.

[26] **Yu, G.**, **Mallat, S. and Bacry, E.** (2008). Audio denoising by time-frequency block thresholding, *Signal Processing, IEEE Transactions on*, *56*(5), 1830–1839.

[27] **Hikichi, T.**, **Delcroix, M. and Miyoshi, M.** (2007). Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations, *EURASIP Journal on Advances in Signal Processing*, *2007*(1), 1–12.

[28] **Miyoshi, M. and Kaneda, Y.** (1988). Inverse filtering of room acoustics, *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *36*(2), 145–152.

[29] **Radlovic, B.D.**, **Williamson, R.C. and Kennedy, R.A.** (2000). Equalization in an acoustic reverberant environment: Robustness results, *Speech and Audio Processing, IEEE Transactions on*, *8*(3), 311–319.

[30] **Mertins, A.**, **Mei, T. and Kallinger, M.** (2010). Room impulse response shortening/reshaping with infinity-and-norm optimization, *Audio, Speech, and Language Processing, IEEE Transactions on*, *18*(2), 249–259.

[31] **Schwartz, O.**, **Gannot, S. and Habets, E.A.** (2015). Multi-microphone speech dereverberation and noise reduction using relative early transfer functions, *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, *23*(2), 240–251.

[32] **Strang, G.** (1986). A proposal for Toeplitz matrix calculations, *Studies in Applied Mathematics*, *74*(2), 171–176.

CURRICULUM VITAE

Name Surname **:** Aziz Koçanaoğulları

Place and Date of Birth **:** İzmir, 05.01.1991

Address **:** İnönü cad. No:374/12
Hatay , İZMİR

E-Mail **:**azizkocana@gmail.com

EDUCATION

B.Sc. **:** Istanbul Technical University (2014)
Electrical and Electronics Faculty
Electronics Engineering

B.Sc. **:** Istanbul Technical University (2015)
Faculty of Science and Letters
Mathematics Engineering

M.Sc. **:** Istanbul Technical University (2016)
Graduate School of Science Engineering and Technology
Telecommunications Engineering