Discriminative Framework for Single Channel Audio Source Separation

Student Name: Arpita Gang

IIIT-D-MTech-ECE

July 8, 2016

Indraprastha Institute of Information Technology
New Delhi

Thesis Advisor
Dr. Pravesh Biyani

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Electronics and Communication Engineering,
with specialization in Communication and Signal Processing

Keywords: Source separation, Discriminative learning, dimension, NMF, dictionary, subspace

Certificate

This is to certify that the thesis titled **"Discriminative Framework for Single Channel Audio Source Separation"** submitted by **Arpita Gang** for the partial fulfilment of the requirements for the degree of *Master of Technology* in *Electronics and Communication & Engineering* is a record of the bonafide work carried out by her under my guidance and supervision at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Dr. Pravesh Biyani
Indraprastha Institute of Information Technology, New Delhi

Abstract

Sound sources are a very common everyday occurrence. But a single audio source is seldom heard alone. There is a sea of applications, like speech recognition, where an isolated sound source is desirable. This makes audio source separation a very important problem. In this thesis, we focus on the single channel source separation (SCSS) problem, which implies separation of individual sources from a single observation. The problem of finding many unknowns from one equation makes this problem ill-posed forcing the use of some prior information for better separation.

Model-based methods for single channel source separation use prior information in the form of learned bases. In case of similar signals like speech, models will be highly overlapping, thus making separation difficult. Thus, the sources should be modeled using proper bases/structure for an effective separation. Along with the model, the parameters of the model also play a vital role in quantifying the quality of separation. In any model, a higher dimension (number of columns) makes it a good fit for the source. But for similar sources, it also makes a good fit for the other source. Thus, dimension of the models are an important factor in deciding the discrimination provided by the models and hence the quality of separation. Also, separating one source at a time from the mixture extricates the problem from balancing the reconstruction of all the sources thus, improving the separation performance.

In this thesis, we propose a novel discriminative learning framework for source separation of audio signals when observed from a single mixture. The framework is generic where we separate one source at a time and embed our dimension search algorithm in the training of discriminative source models. We apply our framework on the NMF based SCSS algorithm. We also propose an alternative structure using dictionary and subspace together for learning source models. We demonstrate a performance improvement in separation for both speech-speech and speech-music mixture.

Acknowledgments

I would like to pay gratitude to everyone who have helped me during the course of my thesis. First and foremost, I am deeply thankful to my advisor Dr. Pravesh Biyani for his continuous guidance and immense support.

I would like to thank my friends, specially Ankita and Suman, for helping me in every small issue be it academic or non-academic. I would also like to thank my friend Himanshi for her continuous support and motivation. Lastly, I acknowledge that without the blessings of my parents, grandparents and the love of my brother I could not have done anything.

Contents

1 Introduction **1**

1.1 Source Separation 2

1.2 Prior work 3

1.3 Motivation and Contribution 5

1.4 Outline 6

2 Single Channel Source Separation **7**

2.1 Overview 7

2.2 Fitting versus Separation 10

2.3 Parameters for quantifying discrimination 10

3 Discriminative Framework for SCSS **13**

3.1 Structures 14

3.1.1 Non-negative Matrix Factorisation 14

3.1.2 Dictionary-Subspace Structure 14

3.2 Proposed Algorithm 15

3.2.1 NMF dictionaries 15

3.2.2 Dictionary-Subspace Structure 16

3.2.3 Discriminative Training: The Framework 17

3.3 Framework for multiple sources 18

3.3.1 Finding source model 19

3.3.2 Finding interferer models 19

4 Results and Discussion **21**

4.1 Dataset 21

4.2 Evaluation Metrics 21

4.3 Results for separation of two sources 22
4.3.1 Parameters used for NMF structure 22
4.3.2 Parameters used for Dictionary-Subspace Structure 22
4.3.3 Performance Evaluation 23
4.4 Results for separation of three sources 26
4.4.1 Parameters used for NMF 27
4.4.2 Parameters for Dictionary-Subspace 27
4.4.3 Performance Evaluation 27

5 Conclusion and Future Work **30**
5.1 Thesis conclusion 30
5.2 Future Scope 30

List of Figures

1.1 General source separation problem 2

4.1 Spectrogram analyis for speech separation: (a)-(b) Original spectrograms (c)-(d) Spectrograms after separation using RNMF (e)-(f)Spectrograms after separation using DF-NMF 24

4.2 Spectrogram analyis for speech-music separation: (a)-(b) Original spectrograms (c)-(d) Spectrograms after separation using RNMF (e)-(f)Spectrograms after separation using DF-NMF 25

4.3 Spectrogram analyis for speech separation: (a)-(c) Original spectrograms (d)-(f) Spectrograms after separation using RNMF (g)-(i)Spectrograms after separation using DF-NMF 28

List of Tables

4.1 Average performance for speech-speech separation 26
4.2 Average performance for speech-music separation 26
4.3 Performance comparison of Dictionary subspace structure 26
4.4 Average performance for separation of three sources 29
4.5 Performance comparison of Dictionary subspace in multiple source scenario 29
4.6 Performance comparsion with NMF-ISS 29

List of abbrevation

ASR	Automatic Speech Regonition
SCSS	Single Channel Source Separation
BSS	Blind Source Separation
ISA	Independent Subspace Analysis
ICA	Independent Component Analysis
NMF	Nonnegative Matrix Factorisation
CASA	Computational Auditory Scene Analysis
CMF	Complex Matrix Factorisation
STFT	Short Time Fourier Transform
DNN	Deep Neural Network
RNN	Recurrent Neural Network
KL	Kullback-Leiber
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
SDR	Source to Distortion Ratio
SIR	Source to Interference Ratio
SAR	Source to Artifacts Ratio

Chapter 1

Introduction

Audio signals are an integral part of day-to-day human life. Various audio signals like speech and music are encountered innumerable times in the course of a day. There are many instances where one comes across a mix of signals like different instruments playing together in an orchestra or people speaking simultaneously etc. Human listeners have little difficulty in concentrating on one person or a particular instrument even when there are several active sources. But recognition of a particular audio signal from a mixture containing multiple sources is rather difficult for computer audition. A common example is the *cocktail party problem*, when a number of speakers are talking simultaneously in the presence of background noises like in a cocktail party and one tries to follow one particular speaker.

The separation of one source from a mixture of sources can have potential applications in many analysis algorithms. Separation of speech signals can help in automatic speech recognition (ASR) [1] and speech coding; separation of musical instruments may be required for music retrieval or music transcription [2]. These are a few examples where source separation is required as a pre-processing step as the ASR or music transcription systems essentially require single source signals to operate on. Other important application areas of source separation include communication systems where separation of two signals can help in mitigation of interference, image processing where separation of an image into texture and cartoon (piece-wise smooth) parts is required for image synthesis and analysis [3] etc.

Source separation is the class of algorithms that deals with this problem. In the past decade, many approaches to solve this problem have been developed. But still, the capacity of machines lie far behind human capability for separation making the source separation problem an open field of research.

1.1 Source Separation

The problem of source separation can be thought of as recovering L sources, $s_l(t), l = 1, 2...L$, from M observations of their mixtures, $y_m(t), m = 1, 2...M$. Figure depicts the general source separation scenario. In a simple case, the source signals can be assumed to arrive at the microphones simultaneously without being filtered i.e., each source contributes to each observed mixture (channel) with a multiplicative gain. Such mixtures are called instantaneous mixtures [1].

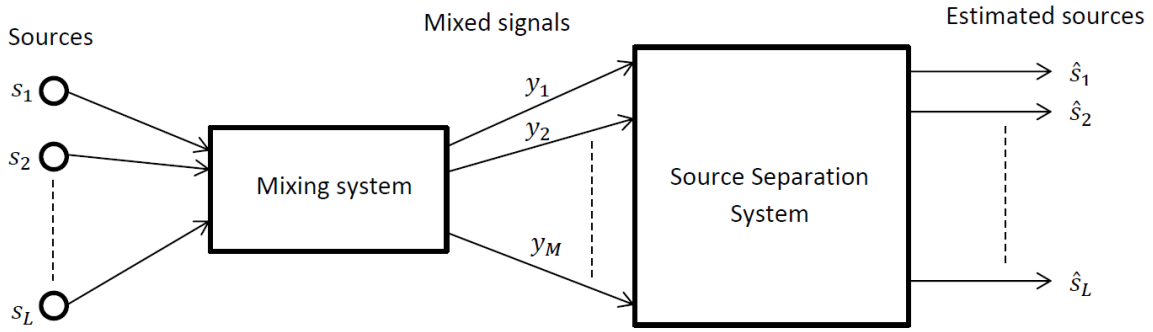


Figure 1.1: General source separation problem

The source mixing system for instantaneous mixtures can be put in vector form as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_M \end{bmatrix} = \begin{bmatrix} a_{11} & .. & .. & .. & a_{1L} \\ a_{21} & a_{22} & .. & .. & a_{2L} \\ .. & .. & .. & .. & .. \\ a_{M1} & .. & .. & .. & a_{ML} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ .. \\ s_L \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{y = As} \tag{1.1}$$

An important issue to be considered here is the relationship between the number of observations M available of a particular mixture and the number of sources L lying in it. When the number of observations is greater than or equal to the number of sources i.e., $M \geq L$, it is equivalent to say that the number of equations is greater than or equal to the number of unknowns, thus making the system overdetermined or critically determined. It is often possible to recover the original sources from the overdetermined system without making any strong prior assumption about the sources. For these noiseless determined (or, overdetermined) cases, there exist a demixing system $W = A^{-1}$ (or, $W = A^{\dagger}$) which when estimated, the sources can be simply recovered as $s = Wy$ [4].

[1] Another type of mixture is convolutive, where the mixtures are addition of filtered sources. This leads to a more difficult problem.

The problem becomes more difficult when the number of observation M is less than number of sources L in which case the system becomes underdetermined. Some strong assumptions and prior information about the sources are required to recover the sources from underdetermined mixtures. Single channel source separation (SCSS) is an extreme case of this underdetermined mixture case wherein only one observation of the mixture of signals is available.

The work in this thesis deals with single channel source separation in the instantaneous mixing case.

1.2 Prior work

Many approaches for solving the SCSS problem have been proposed over the years. As mentioned before, SCSS aims at recovering the underlying sources in a mixture from a single observation. At this point, two cases are possible : first when one has no prior knowledge of the sources and second when some prior information is available. The former scenario is called Blind Source Separation (BSS). Some algorithms have been proposed which operate in a blind manner without making any strong assumptions about the sources except for statistical independence or sparsity. Casey and Westner [5] proposed independent subspace analysis (ISA), an extension of independent component analysis (ICA), for the SCSS problem where basis vectors are learned from the mixed signal spectrogram and are grouped together into subsets, one for each source. ISA does not work well when the underlying sources have overlapping bases. Another widely used method for unsupervised single channel source separation is based on non-negative matrix factorisation (NMF). Many algorithms using NMF in an unsupervised manner have been proposed using sparsity and temporal constraints [6] [7]. It is possible to do the clustering of the bases learned from ISA or NMF based methods in a completely unsupervised manner. But this type of clustering becomes rather difficult when the sources are similar and the bases are overlapping. Computational Auditory Scene Analysis (CASA) is another approach used for such SCSS problems. These methods [8] [9] rely on pitch and harmonic structure of the sources. Similar to above mentioned approaches, CASA methods are not able to segment the mixed signal into the individual sources well enough in all cases, especially in case of similar sources like speech signals.

The second class of algorithms operate with some prior knowledge about the specific sources. Being an underdetermined problem, the SCSS system will have infinite number of solutions. Thus, prior information about the sources can help in finding the actual

solution i.e. in the recovery of the actual sources. The available information regarding the underlying sources are used to learn source models which are representations of the corresponding sources. Separation methods based on source models have been quite successful in the single channel source separation problem. The model-based source separation methods can again be of two types: linear and non-linear. Linear methods are those in which the mixture is expresed as a linear combination of basis vectors representing the sources. Non-linear methods rely on designing more complex structures for separating mixture into the constituents sources.

Linear methods aim at discovering such bases of the sources such that they help in separation. For instance, if two sources lie in $\mathbb{R}^n$ and if the basis from which these sources derive their data points are known and are orthogonal, then one can easily recover these two sources from their mixture by simply projecting the mixture vector on the two orthogonal basis. Jang and Lee [10] propose such a model-based method which separates sources using predefined, source-specific ICA bases learned from training data. Supervised NMF [11] [12] [13] and CMF (Complex Matrix Factorisation) [14] [15] are other widely used approaches in model-based SCSS methods. The work proposed in [11] learns NMF models with a sparsity constraint to help in separation.

Along with being a good representation of the source, the source separation problem demands that the models are able to discriminate between various sources i.e. the learned models are 'discriminative'. Recently, many discriminative learning methods have been shown to produce more effective separation. The method in [12] attempts to learn discriminative models by minimizing the cross-coherence between the basis vectors pertaining to different sources while [13] formulated a training-time optimization of the reconstruction bases using the test-time inference method applied to mixed signals. Another discriminative learning method with NMF is proposed in [16] which optimizes all basis vectors jointly to reconstruct both clean signals and mixed signals well. The work in [17] and [18] also propose discriminative model based methods where the models for the underlying sources are learnt in the form of overcomplete dictionaries. The proposed work in [17] attempts to learn the dictionaries for all the sources simultaneously rather than as independent units and [18] learns a sequence of dictionaries and performs separation in a number of stages. Some non-linear methods based on deep neural networks (DNN) and recurrent neural networks (RNN) have also been proposed recently. The work in [19] propose a joint optimisation of deep learning models with an extra layer of masking. Another method combining NMF and deep network architecture have been proposed in [20] which unfolds the NMF iterations into the layers of the network.

In this thesis, we focus on linear model-based methods for single channel source separation. The work proposed is another step towards discriminative learning of source models.

1.3 Motivation and Contribution

As mentioned above, the SCSS problem requires learning of source models such that they are a good representations of their own sources while being inefficient representations of the other sources lying in the mixture. This is required so that they aid a good separation. In the ideal scenario, mutually orthogonal models for the sources would lead to best separation. But, such models may not exist for sources which are too similar to each other. In such situations, efficient models should be as orthogonal or discriminative as possible to each other. Such models are learned from the training data available for the underlying sources.

Almost all the SCSS methods aim at recovering all the underlying sources simultaneously. A single optimisation problem is solved with the goal of separating and hence reconstructing all the sources at the same time. Thus, a good reconstruction and separation becomes equally important for all the sources. If we can focus on the quality of separation of just one source out of all, a better recovery can be made for it by sacrificing the quality of other sources. To encapsulate this idea, we suggest a solve a separate optimisation problem for each source treating all the other sources as 'interferers'. In effect, we solve as many optimisation problems as the number of components in the mixture. Since every source is separated by solving its own optimisation problem, the reconstruction quality of the corresponding interferers is not relevant. The suggested to-each-its-own framework performs better than a joint separation framework as the later is burdened with the task of balancing the reconstruction of all the components, unlike our proposed framework.

The source models can have different structures like NMF bases, overcomplete dictionaries, subspace etc. Given any source model, the source separation performance also depends on choosing the right parameters like the number of basis vectors of the NMF matrix and dictionary or the sparsity of the dictionary. In a model, a higher number of columns of the model implies a better representation of the given source. But this may also result in the model becoming a better fit for the interferer as well, especially in case of similar sources. Determining an appropriate dimension for the models thus provides another lever for the discriminative source separation. In fact, the similarity between the constituent sources can be different for different mixtures and thus, dimensions should also be chosen specific to the sources in the concerned mixture. We propose to introduce dimension search as

a part of the optimisation problem for discriminative training. To aid the search for dimensions, we introduce certain ratio based parameters that give an idea about the extent of discrimination a given set of models can offer, thus reassuring the quality of separation.

The primary contribution of this thesis is the development of a framework which embodies the following features:

- Solving the separation problem as many times as the number of sources in the mixture so as to recover only one source at a time. We term it as 'to-each its own' framework.
- Introducing dimension search along with the to-each-its-own framework for a better discriminative training.

The framework is generic and can be applied over and above any type of model and can improve its separation performance. We have worked with NMF dictionaries and demonstrated an improvement in performance on using the framework. We also another structure for modeling the sources based on the combined use of dictionary and subspace. The results show an improvement in performance in speech separation and speech-music separation cases.

1.4 Outline

The rest of the thesis is organised as follows: Chapter 2 describes the single channel source separation problem. It also describes the idea for the ratios that can help in the dimension search for the discriminative training of the models. The structures used for modeling the sources i.e. NMF dictionaries and Dictionary-Subspace are explained in Chapter 3. The proposed framework for separation of two sources is also explained here and the extension of the framework for separating more than two sources is presented. Chapter 4 discusses the results for separation of two sources in case of speech-speech and speech-music mixtures. The results for separation of thrre sources in speech-speech case is also presented.

Chapter 2

Single Channel Source Separation

In this chapter, the basic overview of the model-based approach in the single channel source separation case is explained. Along with that, the basic ideas behind the development of the proposed method are presented.

2.1 Overview

The simplest SCSS model is of the form:

$$y(t) = \sum_{l=1}^{L} s_l(t) \tag{2.1}$$

Given $y(t)$, the aim of source separation is to obtain estimates of the L sources, $\hat{s}_l(t), l = 1....L$. The model-based approaches for SCSS work in three stages :

1. **Feature extraction**: The available training data for each source is used to extract features such as Short Time Fourier Transform (STFT) or Mel-features etc.

2. **Training stage**: The extracted features are used to learn models for each source such that they can aid in separation along with reconstruction of the sources.

3. **Separation stage**: Sources are separated by projecting the mixed signal on the learned models with some added constraints to retrieve the individual sources.

In this thesis, we work with STFT features. A discrete-time signal $y(n)$ is multiplied by an analysis window $w(n)$ of length w_{len}. This windowed signal is transformed into

frequency domain by discrete fourier transform (DFT). The resulting spectrum forms one column of the STFT matrix $\underline{\boldsymbol{Y}}$. For the next column, the window is shifted by a certain hop size h. We have used Hamming window as the windowing signal i.e.,

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N}),\ \ 0 \leq n \leq N \tag{2.2}$$

The length of this window w_{len} is $N+1$. For real-valued signals, the DFT results in a complex conjugate symmetric matrix []. Thus, half of the spectrum is dropped for further processing which reduces the lengh of one column to $N = \frac{w_{len}}{2} + 1$. The real-valued time domain addition of eq. 2.1 results in a complex-valued addition in the STFT domain.

$$\underline{\boldsymbol{Y}} = \sum_{l=1}^{L} \underline{\boldsymbol{S_l}} \tag{2.3}$$

The features used to train models are the magnitude spectrogram of the signal obtained as $\boldsymbol{Y}(\omega) = |\underline{\boldsymbol{Y}}(\omega)|$. If all the source signals have equal phase, the magnitude spectrogram of the mixed signal becomes

$$\boldsymbol{Y} = \sum_{l=1}^{L} \boldsymbol{S_l} \tag{2.4}$$

Since, it is unrealistic for all the source signals to have same phase, eq. 2.4 is written as an approximation:

$$\boldsymbol{Y} \approx \sum_{l=1}^{L} \boldsymbol{S_l} \tag{2.5}$$

The magnitude spectrogram of the source signals are used to train models. Most commonly used models in case of single channel separation are NMF bases [12] [13] [11], overcomplete dictionaries [17] [18]. Let the models of the sources $s_1(t), s_2(t)..$ be denoted by $\boldsymbol{D}_1, \boldsymbol{D}_2..$ respectively. The sources are retrieved in the separation stage by solving the following optimisation problem:

$$\underset{\boldsymbol{C}_1, \boldsymbol{C}_2 ... \boldsymbol{C}_L}{\operatorname{argmin}}\ D_\beta(\boldsymbol{Y} \| \sum_{l=1}^{L} \boldsymbol{D}_l \boldsymbol{C}_l)\ \ \text{subject to some constraint on } \boldsymbol{C}_1, \boldsymbol{C}_2 ... \boldsymbol{C}_L \tag{2.6}$$

The equation 2.6 is solved by minmising the β-divergence [21] between $\boldsymbol{Y}$ and $\sum_{l=1}^{L} \boldsymbol{D}_l \boldsymbol{C}_l$. The β-divergence represents different distance measures for different values of β. It is

defined as

$$d_\beta(x\|y) = \begin{cases} \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}) & \beta \in \mathbb{R} \setminus \{0,1\} \\ x\log\frac{x}{y} - x + y & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{2.7}$$

For $\beta = 0$, the metric is equivalent to Itakura-Saito divergence, $\beta = 1$ implies Kullback-Leiber (KL) divergence and $\beta = 2$ represents the squared Euclidean distance. The spectrograms of the underlying sources are then estimated using the model of the individual source and the corresponding co-efficients estimated from 2.6.

$$\hat{\boldsymbol{S}}_l = \boldsymbol{D}_l \boldsymbol{C}_l \qquad \forall l = 1, 2..L \tag{2.8}$$

The final estimates of the source STFT's can be made using either spectral masking or using the phase of mixed signal STFT. The latter is the simplest way of estimating the final STFT using the phase information $\angle\underline{\boldsymbol{Y}}$ of the mixed signal directly.

$$\angle\underline{\boldsymbol{Y}} = \frac{\underline{\boldsymbol{Y}}(\omega)}{\boldsymbol{Y}(\omega)} \qquad\qquad \underline{\hat{\boldsymbol{S}}_l} = \hat{\boldsymbol{S}}_l \angle\underline{\boldsymbol{Y}} \tag{2.9}$$

In the masking method, a spectral mask M_l for each source is built using the estimated spectrograms.

$$\boldsymbol{M}_l = \frac{\hat{\boldsymbol{S}}_l}{\sum_{m=1}^{L} \hat{\boldsymbol{S}}_m} \tag{2.10}$$

Here, the divisions are done element-wise. The final estimate of the source STFT is made using the masks and the mixed signal STFT. This type of signal estimation implies that the separated sources sum up to the mixture.

$$\underline{\hat{\boldsymbol{S}}_l} = \boldsymbol{M}_l \otimes \underline{\boldsymbol{Y}} \tag{2.11}$$

The operator $\otimes$ denotes element-wise multiplication. The estimated source spectrogram is converted back to time domain by applying inverse STFT operation to get a final estimate of the sources $\hat{s}_l(t), l = 1, 2..L$.

$$\hat{s}_l(t) = \text{ISTFT}(\underline{\hat{\boldsymbol{S}}_l}) \tag{2.12}$$

The process of estimation clearly shows that the quality of separation of any source depends on all the models $\boldsymbol{D}_l,\ \forall l = 1, 2..L$.

2.2 Fitting versus Separation

The interpretation of eq. 2.6 is straightforward. Irrespective of the value of β, this separation equation will aim at fitting $\boldsymbol{Y}$ onto $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2 ... \boldsymbol{D}_L]$. Fitting implies any part of the mixed signal spectrogram $\boldsymbol{Y}$ can be represented by any of the models $\boldsymbol{D}_1, \boldsymbol{D}_2 ... \boldsymbol{D}_L$ so long as the total error in reconstruction is minimised. But, separation is a different and more difficult task. A good separation demands each source to be represented dominantly by its own model only. The accuracy of the estimated source $\hat{s}_l(t)$ is related to the efficiency of the model $\boldsymbol{D}_l$ as well the inadequacy of other models in representing $\boldsymbol{S}_l$. This criteria will keep a check on the imperfections in the recovered source. Along with it, the accuracy of estimation is also dependent on the inefficiency of $\boldsymbol{D}_l$ in representing all other sources, which is necessary to restrict interference in $\hat{s}_l(t)$. Thus, the essence of model-based methods lies in an effective training such that the models can differentiate between their own source and other sources during separation.

The level of discrimination that a particular set of models can offer comes from their training. Since every source can have a different degree of similarity with every other source, the level of discrimination required is also different. As described in the previous chapter, number of columns/dimension of the model plays a key role in making the models discriminative. So, the dimension of the models should be adapted according to the sources in the mixture in a way that they provide a good separation. A natural question to ask at this point is: how do we parametrically quantify discrimination? To this end, we define a few ratio based parameters in the next section that naturally quantify the separation that can be achieved using the models obtained from training.

2.3 Parameters for quantifying discrimination

Since we are interested in recovering only one source at a time from a mixture, rest of the constituents are treated as "interferers". To understand the concept of the ratios, let us assume a mixture of two sources. We denote the source to be recovered as $s_s(t)$ and the other source i.e., the interferer as $s_n(t)$. The source and interferer models are denoted by $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$. We denote the spectrograms of training data as $\boldsymbol{S}_s$ and $\boldsymbol{S}_n$. Assuming that the number of frames of the source and interferer available for training are J_s ans J_n respectively, the sizes of $\boldsymbol{S}_s$ and $\boldsymbol{S}_n$ are $N \times J_s$ and $N \times J_n$ where N depends on the FFT size as explained in section 2.1.

When the mixture is composed of only specific source, it should be represented dominantly

by its own model even when offered a concatenated structure $\boldsymbol{D} = [\boldsymbol{D}_s\, \boldsymbol{D}_n]$. In other words, on projection of a source over $\boldsymbol{D}$, the ratio of the energy over its own model to its energy over other source (or interferer) model should be high. Mathematically:

1. Say $\boldsymbol{Y}$ consists of only the source signal $\boldsymbol{S}_s$. On solving (2.6) in this case, the co-efficients obtained are

$$\boldsymbol{C}_{ss}, \boldsymbol{C}_{sn} = \underset{\boldsymbol{C}_{ss}, \boldsymbol{C}_{sn}}{\operatorname{argmin}}\ D_\beta(\boldsymbol{S}_s \| \boldsymbol{D}_s\boldsymbol{C}_{ss} + \boldsymbol{D}_n\boldsymbol{C}_{sn}) \tag{2.13}$$

   Here, $\boldsymbol{C}_{ss}$ and $\boldsymbol{C}_{sn}$ represent the $k_s \times J_s$ and $k_n \times J_s$ co-efficient matrices corresponding to represention of $\boldsymbol{S}_s$ by $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$ respectively. The desired distribution of energy is $E_{ss} = \| \boldsymbol{D}_s\boldsymbol{C}_{ss} \|_F \gg E_{sn} = \| \boldsymbol{D}_n\boldsymbol{C}_{sn} \|_F$ where F denotes the Frobenius norm. Thus, we define a source energy ratio r_s as

$$r_s = \frac{E_{ss}}{E_{sn}} \tag{2.14}$$

2. Similarly, when $\boldsymbol{Y}$ is composed of $\boldsymbol{S}_n$ only, on solving (2.6) the co-efficients are

$$\boldsymbol{C}_{ns}, \boldsymbol{C}_{nn} = \underset{\boldsymbol{C}_{ns}, \boldsymbol{C}_{nn}}{\operatorname{argmin}}\ D_\beta(\boldsymbol{S}_n \| \boldsymbol{D}_s\boldsymbol{C}_{ns} + \boldsymbol{D}_n\boldsymbol{C}_{nn}) \tag{2.15}$$

   $\boldsymbol{C}_{ns}$ and $\boldsymbol{C}_{nn}$ represent the $k_s \times J_n$ and $k_n \times J_n$ co-efficient matrices corresponding to represention of $\boldsymbol{S}_n$ by $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$ respectively. The desired distribution of energy in this case is $E_{nn} = \| \boldsymbol{D}_n\boldsymbol{C}_{nn} \|_F \gg E_{ns} = \| \boldsymbol{D}_s\boldsymbol{C}_{ns} \|_F$. The interferer energy ratio is defined as r_n given by

$$r_n = \frac{E_{nn}}{E_{ns}} \tag{2.16}$$

Clearly, a high value of the source energy ratio r_s would ensure better reconstruction of the source and similarly, a high r_n is required to promote reduced interference in the recovered source.

Additionally, the source model $\boldsymbol{D}_s$ must also be a poor representation of the interferer. This is required to prevent the interferer from getting reconstructed by the source model and hence, to reduce interference in the recovered source. To quantify this, we define an error ratio r_e

$$r_e = \frac{\frac{1}{J_n}\sum_{j=1}^{J_n} \|\boldsymbol{s}_n^j - \boldsymbol{D}_s\boldsymbol{c}_{ns}^j\|_2}{\frac{1}{J_s}\sum_{j=1}^{J_s} \|\boldsymbol{s}_s^j - \boldsymbol{D}_s\boldsymbol{c}_{ss}^j\|_2} \tag{2.17}$$

where, $\boldsymbol{s}_s^j$, $\boldsymbol{s}_n^j$ denote the j^{th} frame (column) of the spectrogram of source and interferer respectively and $\boldsymbol{c}_{ss}^j$, $\boldsymbol{c}_{ns}^j$ are the corresponding co-efficient vectors obtained when using

$\boldsymbol{D}_s$ to reconstruct the frames. Although, a high value of r_n already ensures a good representation of the interferer by $\boldsymbol{D}_n$, having a high value of r_e ensures further reduction of the interference in the source. As we shall see later, we will use r_e to train the source model $\boldsymbol{D}_s$.

The three ratios r_s, r_n and r_e so defined are the parameters that can quantify discrimination. Hence, we use these ratios for performing a dimension search and hence discriminatively training the source and interferer models.

Chapter 3

Discriminative Framework for SCSS

As outlined earlier, the proposed discriminative framework can be applied to any type of models. The framework mainly of two aspects :

- Separating one source at a time treating other as interferes.
- Searching for appropriate dimension of the source and interferer models for proper recovery of the source.

Having described the ratios in the previous chapter, we will utilise them in our framework to train better discriminative models. Specifically, we use the ratio parameters to obtain the source and interferer models such that a certain level of reconstruction accuracy is provided while ensuring that the interference is restricted. Moreover making choice of dimensions as tuneable parameters provides more freedom in training both source and interferer models. It should also be noted that an exhaustive search for all possible values of dimension is computationally infeasible. A search for proper dimensions in joint training i.e., training both the models through a single formulation, would require exploration of a large number of possible combinations. If N is the dimension of the signal space, common values for which are 257 or 513, then each model can have N possible dimensions. This leads to N^L possible combination of dimensions in case of joint training. However, a more efficient way of dimension search can be established when the models are trained separately with the help of the ratios. The proposed algorithm trains the model individually while potentially requiring only a few iterations of the training.

3.1 Structures

The proposed discriminative framework is applied on the NMF dictionaries. Another structure based on the combined use of an overcomplete dictionary and a subspace is also proposed.

3.1.1 Non-negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF) have been used successfully for source separation problems. NMF decomposes a non-negative matrix $\boldsymbol{S} \in \mathbb{R}_{+}^{N \times J}$ into a product of two non-negative matrices $\boldsymbol{D} \in \mathbb{R}_{+}^{N \times k}$ and $\boldsymbol{C} \in \mathbb{R}_{+}^{k \times J}$, where $k \leq N, J$ as follows:

$$\boldsymbol{S} = \boldsymbol{DC} \tag{3.1}$$

The matrices $\boldsymbol{D}$ and $\boldsymbol{C}$ are obtained by solving the following optimisation problem:

$$\boldsymbol{D}, \boldsymbol{C} = \underset{\boldsymbol{D},\boldsymbol{C}}{\operatorname{argmin}} \; D_{\beta}(\boldsymbol{S} \| \boldsymbol{DC}) \;\; \text{subject to } (\boldsymbol{D})_{ij}, (\boldsymbol{C})_{ij} \geq 0 \;\; \forall i, j \tag{3.2}$$

Choosing KL-divergence ($\beta = 1$) as the error metric to be minimized, the solution to 3.2 is obtained by alternating mutiplicative updates as described below:

$$\boldsymbol{D} \leftarrow \boldsymbol{D} \otimes \frac{\frac{\boldsymbol{S}}{\boldsymbol{DC}}\boldsymbol{C}^T}{\mathbf{1}\boldsymbol{C}^T}, \qquad\qquad \boldsymbol{C} \leftarrow \boldsymbol{C} \otimes \frac{\boldsymbol{D}^T\frac{\boldsymbol{S}}{\boldsymbol{DC}}}{\boldsymbol{D}^T\mathbf{1}} \tag{3.3}$$

Here, all the multiplication and division operations are done element-wise. The columns of $\boldsymbol{D}$ are normalised after each iteration. $\mathbf{1}$ is a matrix of ones with size $N \times J$.

3.1.2 Dictionary-Subspace Structure

A dictionary is an overcomplete representation of a class of signals that is used to express any signal pertaining to that class. Due to the overcompleteness, for any given signal, there are many ways to represent it, but normally the sparsest representation is preferred for simplicity and easy interpretability [22]. Specifically, given the training data $\boldsymbol{s}_1, \boldsymbol{s}_2 .... \boldsymbol{s}_J \in \mathbb{R}^N$, learning a dictionary is equivalent to finding a matrix $\boldsymbol{D} \in N \times k$ such that $k > N$

and sparse vectors $\boldsymbol{c}_j$, which can be learned by solving the following optimisation problem:

$$\underset{\boldsymbol{D},\boldsymbol{c}_1 \ldots \boldsymbol{c}_J}{\operatorname{argmin}} \sum_{j=1}^{J} \|\boldsymbol{s}_j - \boldsymbol{D}\boldsymbol{c}_j\|_2^2 + \alpha\|\boldsymbol{c}_j\|_0 \tag{3.4}$$

Due to the non-convexity of the ℓ_0-norm, it is often replaced by the ℓ_1-norm which is its closet convex norm, thus making the optimisation problem of the following form:

$$\underset{\boldsymbol{D},\boldsymbol{c}_1 \ldots \boldsymbol{c}_J}{\operatorname{argmin}} \sum_{j=1}^{J} \|\boldsymbol{s}_j - \boldsymbol{D}\boldsymbol{c}_j\|_2^2 + \alpha\|\boldsymbol{c}_j\|_1 \tag{3.5}$$

On the other hand, a subspace implies a matrix with basis vectors that span the space of a signal. Subspace is, thus, a compact representation of a source with no overcomplete structure. Training a subspace with the training data $\boldsymbol{s}_1, \boldsymbol{s}_2 .... \boldsymbol{s}_J \in \mathbb{R}^N$ is equivalent to finding a $N \times k$ basis matrix such that $k < N$. This problem can be formulated as

$$\underset{\boldsymbol{D},\boldsymbol{c}_1 \ldots \boldsymbol{c}_J}{\operatorname{argmin}} \sum_{j=1}^{J} \|\boldsymbol{s}_j - \boldsymbol{D}\boldsymbol{c}_j\|_2^2 \tag{3.6}$$

We make an attempt to use this difference in representation to our advantage in source separation. Training one of the models as a dictionary and the other as a subspace implies giving one of the sources more freedom to represent itself while restricting the other to a small space.

3.2 Proposed Algorithm

3.2.1 NMF dictionaries

We use NMF dictionaries as in [12] as models for source and interferer. This work trains the two dictionaries jointly adding a regularisation term penalising the cross-coherence between the two dictionaries. Re-writing the optimisation problem of [12] in terms of source and interferer models, we obtain the following

$$\boldsymbol{D}_s, \boldsymbol{D}_n = \underset{\boldsymbol{D}_s,\boldsymbol{D}_n}{\operatorname{argmin}} \; D_{KL}(\boldsymbol{S}_s \| \boldsymbol{D}_s \boldsymbol{C}_{ss}) + D_{KL}(\boldsymbol{S}_n \| \boldsymbol{D}_n \boldsymbol{C}_{nn}) + \lambda \sum_{i,j} (\boldsymbol{D}_s^T \boldsymbol{D}_n)_{ij} \tag{3.7}$$

Unlike [12], we do not train these models jointly. Instead, we break the optimisation problem into two parts. We first train the source NMF dictionary $\boldsymbol{D}_s$ which once trained,

we then follow up with training the interferer dictionary $\boldsymbol{D}_n$.

Optimisation problem for finding the source NMF dictionary $\boldsymbol{D}_s$, for a fixed source dimension k_s is

$$\boldsymbol{D}_s = \underset{\boldsymbol{D}_s, \boldsymbol{C}_{ss}}{\operatorname{argmin}} \; D_{KL}(\boldsymbol{S}_s \| \boldsymbol{D}_s \boldsymbol{C}_{ss}) \tag{3.8}$$

This equation is solved using the multiplicative updates as in 3.3. With the given source NMF dictionary $\boldsymbol{D}_s$, the interferer dictionary is determined $\boldsymbol{D}_n$. For a given value of the dimension k_n of $\boldsymbol{D}_n$, it is obtained by

$$\boldsymbol{D}_n = \underset{\boldsymbol{D}_n, \boldsymbol{C}_{nn}}{\operatorname{argmin}} \; D_{KL}(\boldsymbol{S}_n \| \boldsymbol{D}_n \boldsymbol{C}_{nn}) + \lambda \sum_{i,j} (\boldsymbol{D}_s^T \boldsymbol{D}_n)_{ij} \tag{3.9}$$

Here, λ is the regularisation parameter. (3.9) is solved using multiplicative updates as described below:

$$\boldsymbol{D_n} \leftarrow \boldsymbol{D_n} \otimes \frac{\frac{\boldsymbol{S}_n}{\boldsymbol{D}_n \boldsymbol{C}_{nn}} \boldsymbol{C}_{nn}^T}{\mathbf{1}_n \boldsymbol{C}_{nn}^T + \lambda \boldsymbol{D}_s \bar{\mathbf{1}}_n} \tag{3.10}$$

The columns of $\boldsymbol{D}_n$ are also normalised after each iteration. $\mathbf{1}_n$ and $\bar{\mathbf{1}}_n$ are matrices of ones with size of $\mathbf{1}_n$ being $N \times J_n$ and size of $\bar{\mathbf{1}}_n$ being $k_s \times k_n$. The update of $\boldsymbol{C}_{nn}$ is similar to the update of $\boldsymbol{C}$ as in 3.3. The following problem is then solved for separation using the models $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$.

$$\boldsymbol{C}_s, \boldsymbol{C}_n = \underset{\boldsymbol{C}_s, \boldsymbol{C}_n}{\operatorname{argmin}} \; D_{KL}(\boldsymbol{Y} \| \boldsymbol{D}_s \boldsymbol{C}_s + \boldsymbol{D}_n \boldsymbol{C}_n) \;\; \text{subject to} (\boldsymbol{C}_s)_{ij}, (\boldsymbol{C}_n)_{ij} \geq 0 \tag{3.11}$$

The source $\hat{s}_s(t)$ is then recovered using masking function as described in section 2.1.

3.2.2 Dictionary-Subspace Structure

Our framework aims at recovering only one source at a time. Since we choose to ignore the quality of reconstruction of the interferer in our framework, we model it as a subspace. Similar to NMF dictionaries above, the source model is trained first as an overcomplete dictionary with $k_s > N$.

$$\boldsymbol{D}_s = \underset{\boldsymbol{D}_s, \boldsymbol{C}_{ss}}{\operatorname{argmin}} \; \|\boldsymbol{S}_s - \boldsymbol{D}_s \boldsymbol{C}_{ss}\|_F^2 + \alpha \|\boldsymbol{C}_{ss}\|_1 \tag{3.12}$$

Here, α is the regularisation parameter. This dictionary is learnt using the K-SVD algorithm [23]. Given the source dictionary, the interferer model $\boldsymbol{D}_n$ is trained as a subspace with an incoherence condition similar to the regularisation term used for NMF dictio-

naries. The limited memory BFGS algorithm (L-BFGS) is used to solve the following equation.

$$\boldsymbol{D}_n = \underset{\boldsymbol{D}_n, \boldsymbol{C}_{nn}}{\operatorname{argmin}} \ \|\boldsymbol{S}_n - \boldsymbol{D}_n \boldsymbol{C}_{nn}\|_F^2 + \mu \|\boldsymbol{D}_n^T \boldsymbol{D}_s\|_F^2 \tag{3.13}$$

Having trained both the models $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$, the following separation problem is solved.

$$\boldsymbol{C}_s, \boldsymbol{C}_n = \underset{\boldsymbol{C}_s, \boldsymbol{C}_s}{\operatorname{argmin}} \ \|\boldsymbol{Y} - \boldsymbol{D}_s \boldsymbol{C}_s - \boldsymbol{D}_n \boldsymbol{C}_n\|_F^2 + \alpha \|\boldsymbol{C}_s\|_1 \tag{3.14}$$

The source signal $\hat{s}_s(t)$ is then recovered using the phase of mixed signal.

3.2.3 Discriminative Training: The Framework

We propose to add the optimisation of dimension parameters k_s and k_n for the models of the source and interferer $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$ respectively within our discriminative training framework. The parameters k_s and k_n are searched such that the prefixed ratio parameters r_e, r_s and r_n are satisfied. The search for the dimension k_s of the source dictionary is described in Algorithm 1 . We abbreviate this algorithm for (D)imension (S)earch for (S)ource model as DSS. Searching for an appropriate k_s is essentially finding that value of k_s such that the error ratio r_e in (2.17) gets as close as possible to a predetermined threshold r_{th}. Through experiments we have observed that the ratio r_e is generally monotonically non-decreasing with the dimension variable k_s. Hence a binary search over k_s can be performed. An extremely small value of k_s will quite naturally be inadequate for reconstruction of the source. Also, a high dimension for $\boldsymbol{D}_s$ will tend to make it a good fit for the interferer as well. In such a case, the recovered source will have a high interference even if the error ratio threshold r_{th} is reached. So, the binary search is carried within some bounds $k_{s,min}$ and $k_{s,max}$ for the value of k_s. The value of μ is kept fixed in case of modelling source as a dictionary. For each value of k_s during the search, we solve (3.2) and then calculate r_e. If the threshold is never reached during the binary search, we repeat the search over k_s by lowering the value of the threshold r_{th}. On obtaining an appropriate dimension k_s, the source model $\boldsymbol{D}_s$ is trained using (3.8) in case of NMF structures or (3.12) in case of dictionary-subspace structure.

When the source model $\boldsymbol{D}_s$ is trained, a search for an appropriate dimension of the interferer model $\boldsymbol{D}_n$ is then carried out using the ratios r_s and r_n and we need to ensure both are high. A high value of r_s implies a good reconstruction of the source, while a high value of r_n implies lesser interference. But clearly, there would be a trade-off between these two ratios. To reduce the interference in the source, r_n should be increased but that

would also tend to decrease r_s thereby leading to a poorer reconstruction of the source. So, dimension of the model of interferer $\boldsymbol{D}_n$ is set so that the value of r_n falls within a good range so long as r_s does not fall below a certain minimum value. Algorithm 2 describes the dimension search for the model of the interferer. This algorithm for (D)imension (S)earch for I(N)terferer model is termed as DSN.

Algorithm 1 DSS: Dimension search for source model $\boldsymbol{D}_s$

Input: $\boldsymbol{S}_s$, $\boldsymbol{S}_n$, r_{th}
k_s = NULL;
while k_s == NULL **do**
  Binary search for $r_e = r_{th}$ in $k_{s,min} \leq$ dimension $\leq k_{s,max}$
  if required ratio found **then**
    k_s = dimension obtained from binary search;
  else
    $r_{th} = r_{th} - 0.2$;
  end if
end while
Output: k_s, r_{th}

Algorithm 2 DSN: Dimension search for interferer model $\boldsymbol{D}_n$

Input: $\boldsymbol{S}_s$, $\boldsymbol{S}_n$, $\boldsymbol{D}_s$
$k_n = k_{n,min}$; $in = 1$; {Indicator for search}
while $in == 1$ **do**
  Find $\boldsymbol{D}_n$ with k_n columns
  Find the ratios r_s and r_n
  if $r_s \geq r_{s,min}$ and $r_n \leq r_{n,max}$ **then**
    $k_n = k_n + 5$;
  else
    $in = 0$;
  end if
end while
Output: k_n

The overall algorithm for training the source and interferer models is presented in Algorithm 3 . Once the models are trained, the co-efficient matrices are recovered by solving eq. (3.11) in case of NMF dictionaries or eq. (3.14) in case of dictionary-subspace structure.

3.3 Framework for multiple sources

Due to the advantage of separating one source at a time, our framework can be easily extended for multiple sources. In case of L sources, all models are trained individually

Algorithm 3 Training Algorithm for two sources

1: **Input**: Training data $\boldsymbol{S}_s$, $\boldsymbol{S}_n$
2: $k_s = \text{DSS}(\boldsymbol{S}_s, \boldsymbol{S}_n, r_{th})$.
3: Solve (3.8) or (3.12) to train $\boldsymbol{D}_s$ with k_s columns
4: $k_n = \text{DSN}(\boldsymbol{S}_s, \boldsymbol{S}_n\ \boldsymbol{D}_s)$
5: Solve (3.9) or (3.13) to train $\boldsymbol{D}_n$ with k_n columns.
6: **Output:** $\boldsymbol{D}_s$, $\boldsymbol{D}_n$

while performing dimension search for each model as in previous case. We denote the source model as $\boldsymbol{D}_s$ and the $(L-1)$ interferer models as $\boldsymbol{D}_{n_1}, \boldsymbol{D}_{n_2} ... \boldsymbol{D}_{n_{(L-1)}}$.

3.3.1 Finding source model

Since the degree of similarity is different for every pair of signals, the source signal will have a different error ratio r_e corresponding to each interferer. The interferer which gives the minimum error ratio when represented with the source model has the most similarity with the source. This interferer will tend to create the maximum interference in the source. So, the dimension k_s of the source model $\boldsymbol{D}_s$ is chosen according to the interferer with minimum r_e using the DSS Algorithm 1 . Having found an appropriate dimension for $\boldsymbol{D}_s$, it is then trained using the equation (3.8) for NMF dictionaries or (3.12) in case the source is to be modeled as dictionary.

3.3.2 Finding interferer models

Given the source model $\boldsymbol{D}_s$, the dimension search for the interferer models are carried out using the energy ratios r_s and r_n as in the previous case. As mentioned before, the interferer models are trained individually rather than jointly to ease the dimension search. A question that arises at this point is: Which interferer model to train first? The interferer which gives the maximum r_e is justifiably the one most incoherent with the source. A dimension search for this interferer is carried out with the help of the DSN Algorithm 2. Having found a dimension, the interferer model is then trained according to (3.9) or (3.13). Now, since the quality of separation of the interferers is immaterial, the incoherence between the interferer models can be overlooked. The other interferer models are trained successively according to decreasing order of the error ratio. Discrimination is further promoted in the separation of multiple sources by pushing the interferer models to be close to each other. The formulation for training of the successive interferer models is explained in below.

- NMF:

$$\boldsymbol{D}_{n_l} = \underset{\boldsymbol{D}_{n_l}, \boldsymbol{C}_{nn_l}}{\operatorname{argmin}} \; D_{KL}(\boldsymbol{S}_{n_l} \| \boldsymbol{D}_{n_l} \boldsymbol{C}_{nn_l}) + \lambda_1 \sum_{i,j} (\boldsymbol{D}_s^T \boldsymbol{D}_{n_l})_{ij} + \lambda_2 \sum_{q=1}^{l-1} \sum_{i,j} (\boldsymbol{D}_{n_l} - \boldsymbol{P}_q \boldsymbol{D}_{n_l})_{ij} \quad \forall l = 2, 3...(L-1) \tag{3.15}$$

  $\boldsymbol{P}_q$ is the projection operator of the space of q_{th} interferer model.

- Dictionary-Subspace:

$$\boldsymbol{D}_{n_l} = \underset{\boldsymbol{D}_{n_l}, \boldsymbol{C}_{nn_l}}{\operatorname{argmin}} \; \|\boldsymbol{S}_{n_l} - \boldsymbol{D}_{n_l} \boldsymbol{C}_{nn_l}\|_F^2 + \mu_1 \|\boldsymbol{D}_{n_l}^T \boldsymbol{D}_s\|_F^2 + \mu_2 \sum_{q=1}^{l-1} \|\boldsymbol{D}_{n_l} - \boldsymbol{P}_q \boldsymbol{D}_{n_l}\|_F^2 \tag{3.16}$$

The formal description for training the source and interferer models in case of separation of multiple sources is described in Algorithm 4 .

Algorithm 4 Training Algorithm for multiple sources

Input: $\boldsymbol{S}_s$, $\boldsymbol{S}_{n_1}$,... $\boldsymbol{S}_{n_{(L-1)}}$
for $l = 1$ to $(L-1)$ **do**
    $[r_e(l), d(l)] = \text{DSS}(\boldsymbol{S}_s, \boldsymbol{S}_{n_l}, r_{th})$
end for
$[r_{sort}, \text{ind}] = \text{sort}(r_e)$; {sort in descending order}
$k_s = d(ind(L-1))$;
Train $\boldsymbol{D}_s$ with k_s columns according to (3.8) or (3.12)
$c = ind(1)$;
$k_{n_c} = \text{DSN}(\boldsymbol{S}_s, \boldsymbol{S}_{n_c}, \boldsymbol{D}_s)$
Train $\boldsymbol{D}_{n_c}$ with k_{n_c} columns by solving (3.9) or (3.13)
for $l = 2$ to $(L-1)$ **do**
    $c = ind(l)$;
    $k_{n_c} = \text{DSN}(\boldsymbol{S}_s, \boldsymbol{S}_{n_c}, \boldsymbol{D}_s)$
    Train $\boldsymbol{D}_{n_c}$ with k_{n_c} columns using (3.15) or (3.16).
end for
Output: $\boldsymbol{D}_s$, $\boldsymbol{D}_{n_1}$,...,$\boldsymbol{D}_{n_{(L-1)}}$

On obtaining all the individual interferer models, the accumulated interferer model is obtained as $\boldsymbol{D}_n = [\boldsymbol{D}_{n_1} \boldsymbol{D}_{n_2} ... \boldsymbol{D}_{n_{(L-1)}}]$. The co-efficient matrices are then recovered as in case of two sources, from which the final source signal $\hat{s}_s(t)$ is estimated.

Chapter 4

Results and Discussion

4.1 Dataset

The algorithm was tested for separation of two speech signals and speech-music signals. For the speech case, the algorithm was evaluated on a total of 20 speakers (10 male and 10 female) taken from the TIMIT 16k database [24] which has 10 sentences per speaker. 9 sentences were used for training and one was for testing. The music data was taken from the piano society website [25]. Around 1.5 minutes of data is used for training and another clip from the same artist was used for testing. The mixed signal was formed by adding two signals at a signal to signal ratio of 0 dB. Framing of the signals was done using a Hamming window of length 512 with 75% overlap and a 512 point FFT was taken.

4.2 Evaluation Metrics

The efficacy of separation is measured using the Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and Signal to Artifacts Ratio (SAR). Let $\hat{s}_j(t)$ be the j^{th} estimated source. Let $\Pi\{y_1,y_2...y_k\}$ denote the orthogonal projector onto the space spanned by the vectors $y_1,y_2...y_k$. The projector is a $T \times T$ matrix, where T is the length of these vectors [4]. Two orthogonal projectors are considered as:

$$P_{s_j} = \Pi\{s_j\} \tag{4.1}$$

$$P_s = \Pi\{(s_{j'})_{1\leq j'\leq L}\} \tag{4.2}$$

Then $\mathrm{s}_{\mathrm{target}}$, $\mathrm{e}_{\mathrm{interf}}$ and $\mathrm{e}_{\mathrm{artif}}$ are defined by:

$$\mathrm{s}_{\mathrm{target}} = P_{s_j}\hat{s}_j \tag{4.3}$$

$$\mathrm{e}_{\mathrm{interf}} = P_s\hat{s}_j - P_{s_j}\hat{s}_j \tag{4.4}$$

$$\mathrm{e}_{\mathrm{artif}} = \hat{s}_j - P_s\hat{s}_j \tag{4.5}$$

SDR, SIR and SAR are thus defined as follows:

$$\mathrm{SDR} = 10\log_{10}\frac{\|\mathrm{s}_{\mathrm{target}}\|^2}{\|\mathrm{e}_{\mathrm{interf}}+\mathrm{e}_{\mathrm{artif}}\|^2} \tag{4.6}$$

$$\mathrm{SIR} = 10\log_{10}\frac{\|\mathrm{s}_{\mathrm{target}}\|^2}{\|\mathrm{e}_{\mathrm{interf}}\|^2} \tag{4.7}$$

$$\mathrm{SAR} = 10\log_{10}\frac{\|\mathrm{s}_{\mathrm{target}}+\mathrm{e}_{\mathrm{interf}}\|^2}{\|\mathrm{e}_{\mathrm{artif}}\|^2} \tag{4.8}$$

The BSS evaluation toolbox [26] was used for calculation of the evaluation metrics.

4.3 Results for separation of two sources

A small dimension for either of the models $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$ will hamper the recovery of the source signal. k_s when kept too small will lead to a poor reconstruction of the source and a small value for k_n will lead to more interference in the source and so, small values for both the models are avoided.

4.3.1 Parameters used for NMF structure

The values of $k_{s,min}$ and $k_{n,min}$ are fixed to be 15 while $k_{s,max}$ is taken to be 60. Experiments have shown that $r_{th} \approxeq 3$ is a good value while separating speech files. Also, r_e attains higher values while separation of speech and music files and so the threshold was fixed at 6 in speech-music case. The value of $r_{s,min}$ is fixed at 4 and the value of $r_{n,max}$ is 30 respectively. The value of λ is chosen to be 100.

4.3.2 Parameters used for Dictionary-Subspace Structure

The values of $k_{s,min}$ and $k_{n,min}$ are fixed to be 300 while $k_{s,max}$ is taken to be 750 for the dictionary structure. The value of r_{th} is kept same as the NMF case. The value of $r_{s,min}$

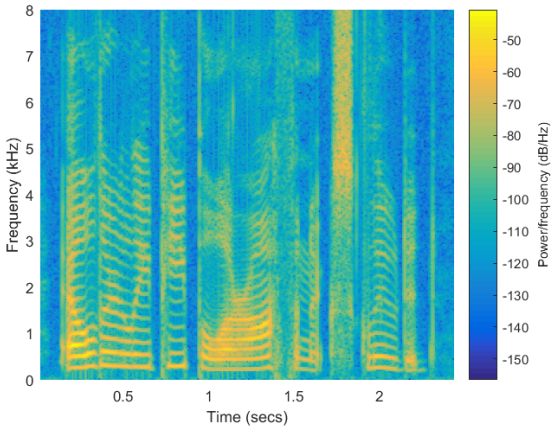
is fixed at 6 and the value of $r_{n,max}$ is 30 respectively. The value of α and μ is chosen to be 0.1 and 1 respectively.

4.3.3 Performance Evaluation

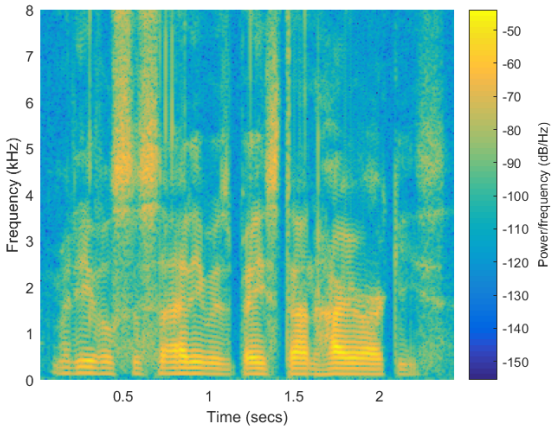
Our experiments show that incorporation of dimension search in the training process and separation of individual sources significantly improves the separation for both the sources. To demonsrate the effect of dimension search, we present the spectrographic analysis of a test case. Figure 4.1 shows the analysis for the separation of two speech signals. The original spectrograms are shown in Figures 4.1a and 4.1b. We applied our framework on NMF dictionaries as in [12], referred to as RNMF which solves (3.7). The spectrograms in figure 4.1c and 4.1d show the separation achieved using the RNMF. Application of our discriminative framework over the RNMF, which we here call DF-NMF, gives the separation as depicted in figures 4.1e and 4.1f. Figure 4.2 shows the spectrograms for separation of a speech and piano signal. It is visibly clear that a simple search for dimension and separating one source at a time gives a better separation.

Along with RNMF, we also compare our results with the dictionary learning based approach in [18], called SDDL, which extracts the sources in a number of levels to achieve better separation and with the joint training formulation proposed in [17]. For speech-speech case, a total of 18 trials were performed, 6 for each case: F+M, F+F and M+M where F refers to female and M refers to male speaker. In case of speech-music, 10 speakers including male and female were used. Table 4.1 compare the average performance of the algorithms in case of separation of speech signals from two sources. The results show that our approach is able to achieve better SDR and SIR compared to RNMF and SDDL while the SAR is lower indicating that our framework introduces a little more artifacts at the cost of lesser interference and distortion. The comparsion of our method with DDL shows that the separation quality of both the approaches is similar. But the time complexity of DDL is much higher than DF-NMF. Table 4.2 depicts the average performance of the algorithms for the separation of speech and music signals. The use of framework lowers the SIR of speech signals as compared to RNMF. But the low SAR for speech sources and low SIR for music sources indicate that reconstructed music signal is composed of high proportion of music as well as high proportion of the speech signal defeating the cause of separation. Overall SDR is better.

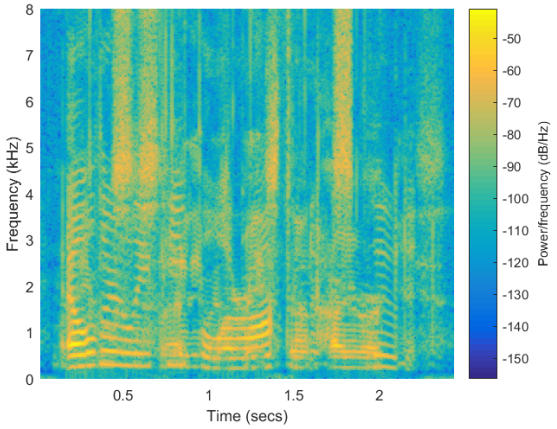
The framework on the alternative structure of dictionary-subspace is referred to as DF-DS in this work. It was observed that separation obtained from DF-DS was not upto the mark as compared to other methods. Although the interference introduced in the



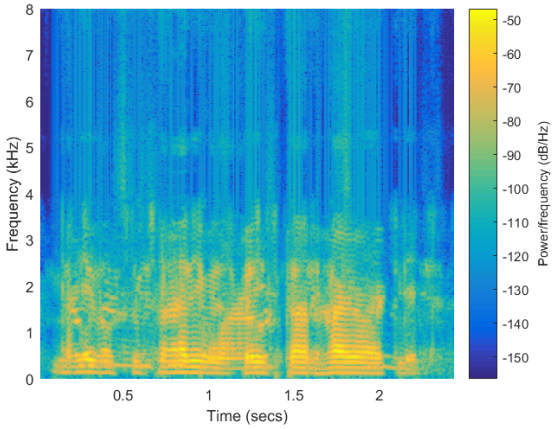
(a) Source 1: Original Spectrogram



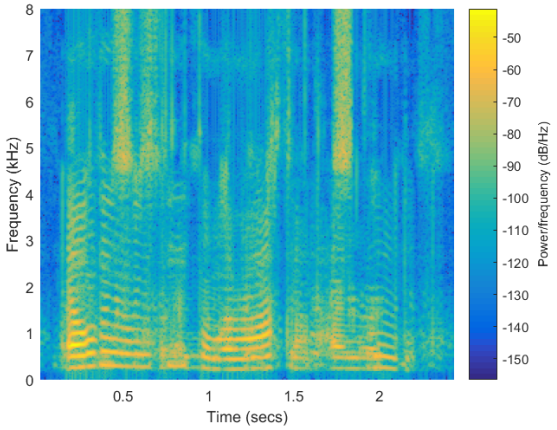
(b) Source 2: Original Spectrogram



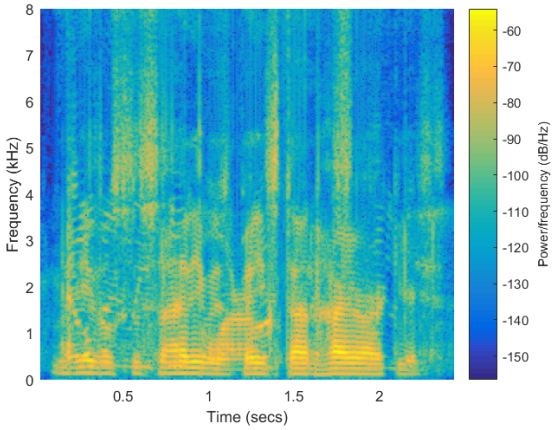
(c) Source 1: After separation using RNMF



(d) Source 2: After separation using RNMF

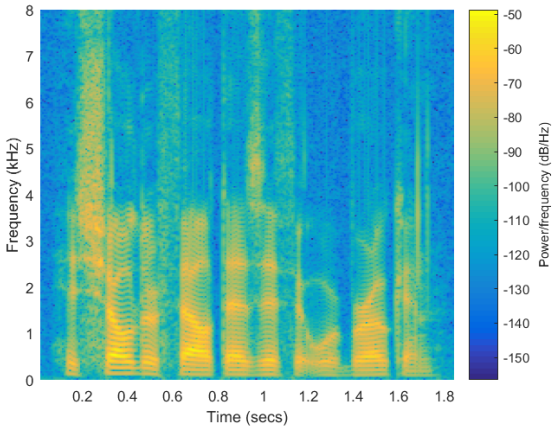


(e) Source 1: After separation using DF-NMF

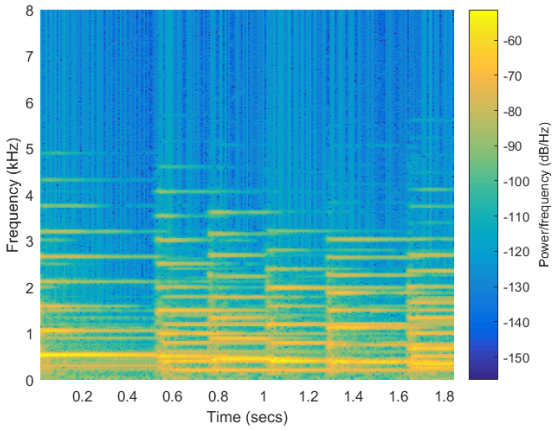


(f) Source 2: After separation using DF-NMF

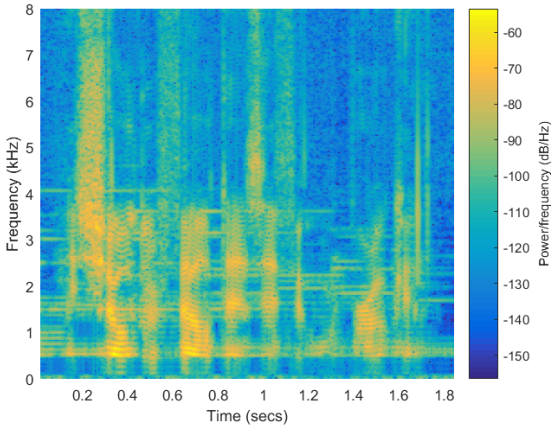
Figure 4.1: Spectrogram analyis for speech separation: (a)-(b) Original spectrograms (c)-(d) Spectrograms after separation using RNMF (e)-(f)Spectrograms after separation using DF-NMF



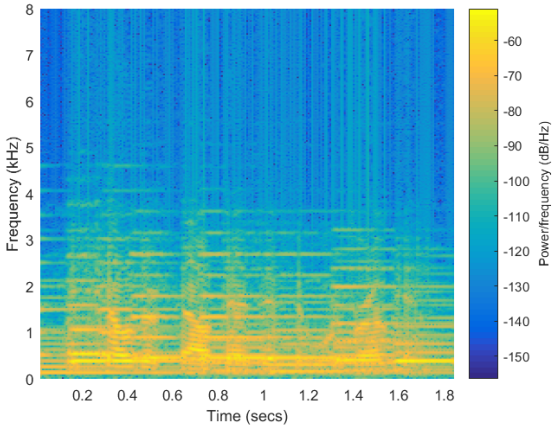
(a) Speech: Original Spectrogram



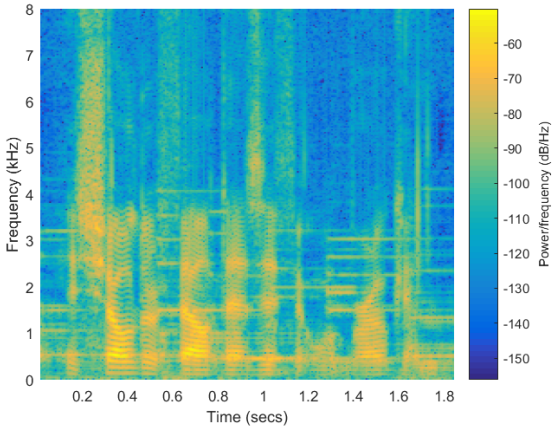
(b) Music: Original Spectrogram



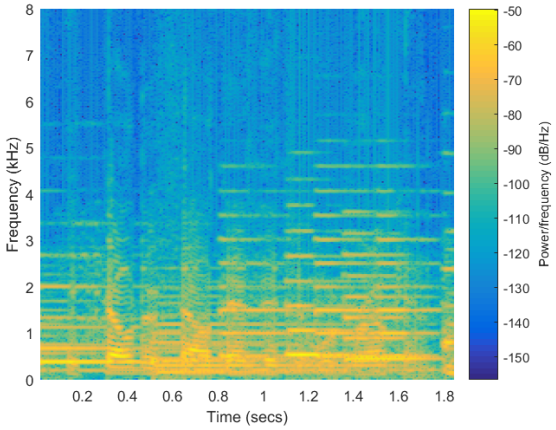
(c) Speech: After separation using RNMF



(d) Music: After separation using RNMF



(e) Speech: After separation using DF-NMF



(f) Music: After separation using DF-NMF

Figure 4.2: Spectrogram analyis for speech-music separation: (a)-(b) Original spectrograms (c)-(d) Spectrograms after separation using RNMF (e)-(f)Spectrograms after separation using DF-NMF

recovered source was less, the SAR was significantly low making the overall distortion high. Five single channel speech mixtures were separated using the dictionary subspace structure and a fall in SAR was seen in every case. Table 4.3 shows the comparison.

		DF-NMF	RNMF [12]	SDDL [18]	DDL [17]
F+M	SDR	6.46	5.6	5.82	6.3
	SIR	8.57	7.01	8.49	8.62
	SAR	11.43	12.32	10.01	10.99
F+F	SDR	4.52	3.78	2.38	3.53
	SIR	6.45	4.95	5.31	5.43
	SAR	10.29	12.29	7.1	9.75
M+M	SDR	3.95	3.56	2.17	2.21
	SIR	5.82	4.65	4.68	5.15
	SAR	10.87	12.35	7.46	7.13

Table 4.1: Average performance for speech-speech separation

		DF-NMF	RNMF [12]	SDDL [18]	DDL [17]
Speech	SDR	7.32	6.2	3.06	4.7
	SIR	10.23	13.88	6.05	7.09
	SAR	11.14	7.34	7.64	10.27
Music	SDR	5.04	2.78	2.85	4.05
	SIR	6.66	3.2	5.53	7.08
	SAR	11.33	14.95	7.5	8.32

Table 4.2: Average performance for speech-music separation

	DF-DS	DF-NMF	RNMF [12]	SDDL [18]	DDL [17]
SDR	4.15	6.18	5.5	5.52	6.06
SIR	10.42	8.27	6.69	8.42	8.37
SAR	6.02	11.52	13.05	9.61	10.87

Table 4.3: Performance comparison of Dictionary subspace structure

4.4 Results for separation of three sources

For the multiple source scenario, mixtures consisting of three speech sources were tested.

4.4.1 Parameters used for NMF

The threshold r_{th} is kept same as the two source case i.e $r_{th} = 3$. The value of $k_{s,min}$ and $k_{n,min}$ was fixed at 15 and $k_{n,max}$ was taken to be 45. The values of λ_1 and λ_2 are 100 and 10 respectively.

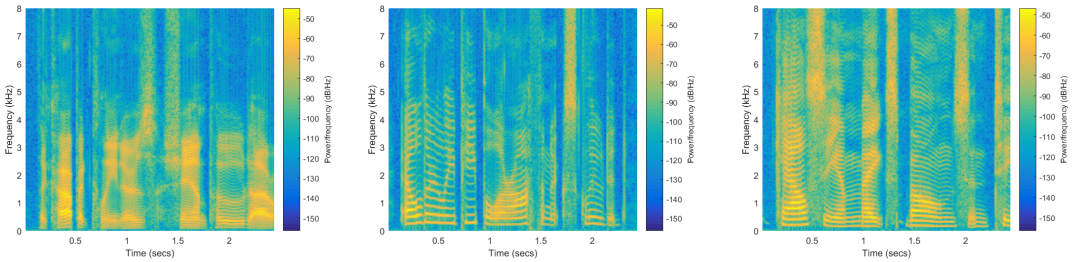
4.4.2 Parameters for Dictionary-Subspace

The threshold r_{th} is kept same as the two source case i.e $r_{th} = 3$. The value of $k_{s,min}$ and $k_{n,min}$ was fixed at 300 and $k_{n,max}$ was taken to be 600. The values of μ_1 and μ_2 are 1 and 10 respectively.

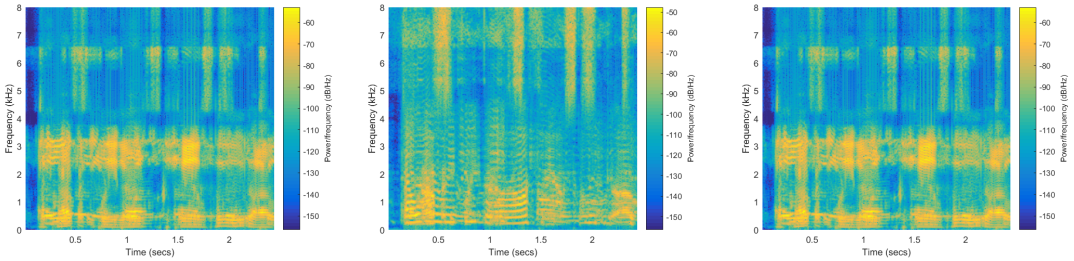
4.4.3 Performance Evaluation

The incorporation of dimension search and separation of one source at a time gives a significant improvement in the three source separation as compared to other methods. Figure ‘4.3 shows the spectrogram analysis fr separation of three sources. We extend the formulation of RNMF presented in [12] for separation of three sources. We compare our proposed method with the RNMF method extended for three sources.

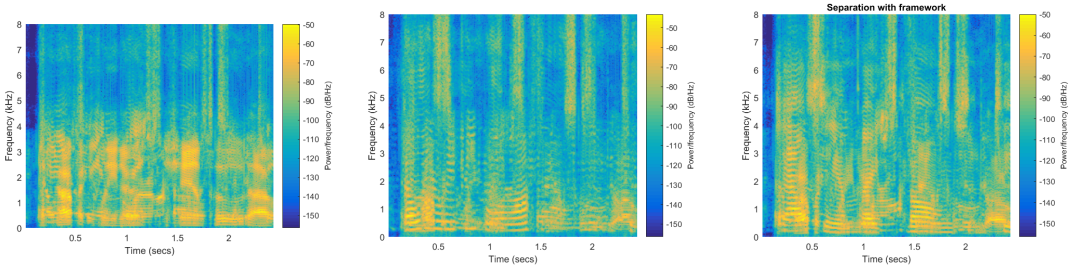
The method of DDL proposed in [17] is applicable for the general case of more than 2 sources. So, a comparison with DDL is also presented. A total of 32 trials were performed: 8 for each case of 1M + 2F, 1F + 2M, 3F and 3M. Table 4.4 shows a comparison of the average performance of the algorithms. The table clearly shows that the proposed method outperforms the other approaches in multiple source scenario. Similar to the case of two sources, the separation using dictionary subspace shows a decrease in SAR. The averag performance comparison of DF-DS with other methods for separation of five mixtures is shown in Table. Another method for separation of multiple sources is proposed in [27] which combines source separation with source coding. This method is referred to as NMF-ISS. This is a semi-blind approach which uses NMF bases learned from source spectrograms in a coded form to initialise the bases and gain matrices for decomposing the mixed signal feature matrix. This method used Mel-features and requires equal length of test and training data. For making a comparison with NMF-ISS, we used half of the available data i.e., 5 sentences per speaker for training and generated a single test file with the remaining 5 sentences. 12 single channel mixtures of three speech sources were separated in this setting. The comparison of DF-NMF and NMF-ISS is shown in Table 4.6



(a) Source 1: Original Spectrogram (b) Source 2: Original Spectrogram (c) Source 3: Original Spectrogram



(d) Source 1: After separation using RNMF (e) Source 2: After separation using RNMF (f) Source 3: After separation using RNMF



(g) Source 1: After separation using DF-NMF (h) Source 2: After separation using DF-NMF (i) Source 3: After separation using DF-NMF

Figure 4.3: Spectrogram analyis for speech separation: (a)-(c) Original spectrograms (d)-(f) Spectrograms after separation using RNMF (g)-(i)Spectrograms after separation using DF-NMF

		DF-NMF	RNMF [12]	DDL [17]
1M+2F	SDR	2.28	0.45	-2.39
	SIR	4.57	2.5	1.32
	SAR	7.98	8.03	3.51
1F+2M	SDR	1.74	0.88	-1.09
	SIR	3.6	2.84	2.17
	SAR	8.54	8.02	4.47
3F	SDR	0.78	-0.39	-10.12
	SIR	2.29	0.98	-0.87
	SAR	9.11	8.5	-4.69
3M	SDR	0.34	-0.018	-4.32
	SIR	1.71	1.87	-0.038
	SAR	8.74	7.65	2.31

Table 4.4: Average performance for separation of three sources

	DF-DS	DF-NMF	RNMF [12]	DDL [17]
SDR	-1.51	2.51	0.62	-1.19
SIR	3.02	4.62	2.75	2.09
SAR	2.49	8.81	8.47	4.61

Table 4.5: Performance comparison of Dictionary subspace in multiple source scenario

	DF-NMF	NMF-ISS [27]
SDR	-0.137	-4.35
SIR	2.61	-1.61
SAR	5.45	3.55

Table 4.6: Performance comparsion with NMF-ISS

Chapter 5

Conclusion and Future Work

5.1 Thesis conclusion

In this thesis, we present a novel framework for discriminative training of models for single channel source separation problem. The proposed framework embodies the concept of searching for an appropriate dimension of the models while solving an individual optimisation problem for every source in the mixture treating other sources as interferers. The framework uses the concept of certain ratios that quanitfy discrimination and thus help in training source and interferer models better suited for separation of the source.

Our proposed method is generic and can be applied over any model. We have used the framework on NMF dictionaries and showed that simply using the framework on existing model improves the separation performance. W also applied on framework on a new structure of dictionary-subspace and demonstrated its performance via simulations.

5.2 Future Scope

Our approach also opens up the possibility of finding theoretical guarantees in source separation. The notion of ratios can also be applied on non-linear separation methods like neural networks. Thus, using the framework for better training of the networks remains as a work for future.

Bibliography

[1] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 826–835, April 2014.

[2] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2096–2107, Oct 2013.

[3] J. Bobin, J.-L. Starck, J. Fadili, Y. Moudden, and D. Donoho, "Morphological component analysis: An adaptive thresholding strategy," *Image Processing, IEEE Transactions on*, vol. 16, no. 11, pp. 2675–2681, Nov 2007.

[4] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[5] M. Casey and W. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proceedings of the International Computer Music Conference, Berlin*, 2000.

[6] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[7] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*. Springer, 2004, pp. 494–499.

[8] J. Nix, M. Kleinschmidt, and V. Hohmann, "Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction," in *In EUROSPEECH*, 2003, pp. 1441–1444.

[9] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *The Journal of the Acoustical Society of America*, vol. 60, no. 4, 1976.

[10] G. jin Jang, T. won Lee, T. won Lee, J. franois Cardoso, E. Oja, and S. ichi Amari, "A maximum likelihood approach to single-channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365–1392, 2003.

[11] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *in International Conference on Spoken Language Processing (INTERSPEECH)*, 2006.

[12] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation." in *INTERSPEECH*, 2013, pp. 808–812.

[13] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Proc. of ISCA Interspeech*, Sep. 2014.

[14] B. J. King and L. Atlas, "Single-channel source separation using complex matrix factorization," *Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 8, pp. 2591–2597, Nov. 2011. [Online]. Available: http://dx.doi.org/10.1109/TASL.2011.2156786

[15] B. King and L. E. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 4206–4209.

[16] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3749–3753.

[17] G. Bao, Y. Xu, and Z. Ye, "Learning a discriminative dictionary for single-channel speech separation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 7, pp. 1130–1138, July 2014.

[18] Y. Xu, G. Bao, X. Xu, and Z. Ye, "Single-channel speech separation using sequential discriminative dictionary learning," *Signal Processing*, vol. 106, pp. 134 – 140, 2015.

[19] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 1562–1566.

[20] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep nmf for speech separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 66–70.

[21] N. L. H. M. C. J. Ayanendranath Basu, Ian R. Harris, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998. [Online]. Available: http://www.jstor.org/stable/2337385

[22] G. Chen and D. Needell, "Compressed sensing and dictionary learning," *Preprint*, vol. 106, 2015.

[23] M. Aharon, M. Elad, and A. Bruckstein, "k -svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[24] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351 – 356, 1990.

[25] http://www.pianosociety.com, 2009.

[26] C. Fevotte, R. Gribonval, and E. Vincent, "Bss eval toolbox user guide," *IRISA, Rennes, France, Tech. Rep*, 2005, [Online]: Available http://www.irisa.fr/metiss/bss eval/.

[27] C. Rohlfing, J. M. Becker, and M. Wien, "Nmf-based informed source separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 474–478.