INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

On Discriminative Framework for Single Channel Audio Source Separation

Arpita Gang, Pravesh Biyani

IIIT Delhi, India
arpita1478@iiitd.ac.in, praveshb@iiitd.ac.in

Abstract

Single channel source separation (SCSS) algorithms that utilise discriminative source models perform better in comparison to those that are trained independently. However, all the aspects of training discriminative models have not been addressed in the literature. For instance, the choice of dimensions of source models (number of columns of NMF, Dictionary etc) not only influences the fidelity of a given source but also impacts the interference introduced in it. Therefore choosing a right dimension parameter for every source model is crucial for an effective 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. Further, separation of a given constituent from a mixture, assuming remaining to be interferers, offers more freedom for the particular constituent and hence provide better separation. In this paper, we propose a generic discriminative learning framework 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 algorithms and demonstrate a performance improvement in separation for both speech-speech and speech-music mixture.

Index Terms: discriminative training, source separation, NMF

1. Introduction

Source separation involves separating the constituent sources or components from their mixture. A classic example of the source separation problem is the *cocktail party problem* where a listener tries to follow one voice in a mix of several people talking together. The extraction of the component signals from the mixed signal can have many potential applications: separation of two speech signals can help in automatic speech recognition (ASR) [1]; separation of vocal and music may be required for music retrieval [2]; in case of communication systems, separation of two different signals will help mitigate interference; separation of an image into texture and cartoon (piecewise smooth) parts for image analysis or synthesis [3].

This paper focusses on the *single channel source separation* (SCSS) problem when only a single observation of the mixture of two signals is available, even though our method is also applicable for a mixture of more than two sources. The SCSS system is underdetermined implying infinite possibilities of set of sources forming a given observed mixture. Recovery of the actual constituents of a given mixture thus requires prior information about its constituents. For instance, when the 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. Single channel source separation (SCSS) problem thus is about discovering structures that not only represents each of the constituent sources but that also aids in the separation of these sources from a mixture. These structures can be derived from source-specific training data in the form of learned source models. Ideally, mutually orthogonal source models that also represent the sources accurately would lead to exact recovery in a noiseless mixing case. Generally such models may not exist as the sources may be too "similar" to each other, e.g, speech of two males or females. Therefore one would like to learn models which are as orthogonal as possible or, in other words, source models should be sufficiently "discriminative". Once the models are learned, roughly speaking, the mixture is projected on these models by solving an optimisation problem to recover/reconstruct the individual components.

1.1. Related work

Various work on the SCSS problem have been proposed using learned models. Casey and Westner [4] proposed independent subspace analysis (ISA), an extension of independent component analysis (ICA) where basis vectors are learned from the mixed signal spectrogram and are grouped using a cross-entropy matrix. ISA does not work well when the underlying sources have overlapping bases. Non-negative matrix factorization (NMF) [5][6][7] and complex matrix factorization (CMF) [8][9] are other widely used approaches for SCSS problem. Authors in [5] attempts to solve this problem by using temporal information while [6] uses sparsity along with the non-negativity constraints. A discriminative approach for NMF is proposed in [7] which attempts to minimize the cross-coherence between the basis vectors pertaining to different sources and in [10] where the test-time objective for separation is incorporated in the formulation. The work in [11] and [12] also propose discriminative model based method where the models for the underlying sources are learnt in the form of overcomplete dictionaries. The proposed work in [11] attempts to learn the dictionaries for all the sources simultaneously rather than as independent units and [12] learns a sequence of dictionaries and performs separation in few stages.

1.2. Motivation and contribution

Almost all of the SCSS work has used a joint separation formulation, wherein the two (or more) sources are separated (and reconstructed) by solving a single optimisation problem. In our work, we suggest a separate optimisation formulation for every source (or component), where to separate one component, the other component is treated as an interferer. Therefore, for recovery of each component we train two models, one for the source and other for the interferer. 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

http://dx.doi.org/10.21437/Interspeech.2016-701

interferer is not relevant. In a way the quality of the separation of one source can be improved at the cost of the other source which acts as an interferer. 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 framework.

Given source models like NMF, dictionary, subspace etc, 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 the NMF based source model a higher number of columns of the model (NMF basis matrix) 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. Determining an appropriate dimension for the source model thus provides another lever for the discriminative source separation. Not only the source model, but the efficacy of the above discussed to-each-its-own framework is also dependant on the choice of dimension of the interferer model. To the best of our knowledge, the current literature does not discuss dimension parameter optimisation as part of discriminative learning.

Our framework is general enough and can be easily applied over and above the discriminative NMF, dictionary or subspace based source models and can improve their separation performance. For this paper, we have worked with NMF dictionaries as in [7] and have improved their performance by determining better choices for the number of basis vectors and by eventually separating one source at a time. We show through simulations that our framework on the NMF based source separation in [7] outperforms other approaches like [11] and [12] for both speech-speech and speech-music separation tasks.

2. Single channel source separation

In single channel source separation problems, it is required to recover the underlying signals from one observation i.e., given

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

the estimates of the L signals, $\hat{s}_l(t), l = 1....L$ are to be found. We focus specifically on the case when $L = 2$. A Short Time Fourier Transform (STFT) of overlapping frames is applied to the audio signals which results in the corresponding matrix, each frame being a column. Let the STFT matrix of the mixed signal be denoted by $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{S}}_1$, $\tilde{\boldsymbol{S}}_2$ be the STFT matrices of the underlying signals $s_1(t)$, $s_2(t)$. The relation between the STFT matrices is

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{S}}_1 + \tilde{\boldsymbol{S}}_2 \tag{2}$$

The magnitude of this matrix represents the spectrogram of the corresponding signal. The spectrograms so formed are denoted by uppercase bold letters. Ignoring the phase of the signals, the spectrogram of the mixed signal is approximated as follows:

$$\boldsymbol{Y} \approx \boldsymbol{S}_1 + \boldsymbol{S}_2 \tag{3}$$

We first explain the NMF [7] methodology of separation of sources. Let the NMF dictionaries of the two sources be $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$. Choosing KL-divergence as the error metric to be minimised, the NMF framework attempts to separate the sources by solving the following optimisation problem:

$$\boldsymbol{C}_1, \boldsymbol{C}_2 = \underset{\boldsymbol{C}_1,\boldsymbol{C}_2}{\text{argmin}} D_{KL}(\boldsymbol{Y} \| \sum_{l=1,2} \boldsymbol{D}_l \boldsymbol{C}_l) \tag{4}$$
$$\text{subject to } (\boldsymbol{D}_l)_{ij}, (\boldsymbol{C}_l)_{ij} \geq 0 \;\; \forall i,j \text{ and } l = 1,2$$

The trained models and the estimated coefficient matrix give the initial estimates of the spectrograms of the two sources.

$$\hat{\boldsymbol{S}}_{1,init} = \boldsymbol{D}_1\boldsymbol{C}_1, \qquad \hat{\boldsymbol{S}}_{2,init} = \boldsymbol{D}_2\boldsymbol{C}_2 \tag{5}$$

These initial estimates of the spectrograms are used to build spectral masks from which final estimates of the sources $\hat{s}_l(t), l = 1, 2$ are made as follows:

$$\hat{s}_l(t) = \mathcal{F}^{-1}\{\boldsymbol{M}_l \otimes \tilde{\boldsymbol{Y}}\} \quad \boldsymbol{M}_l = \frac{\hat{\boldsymbol{S}}_{l,init}}{\hat{\boldsymbol{S}}_{l,init} + \hat{\boldsymbol{S}}_{m,init}}, l \neq m \tag{6}$$

The multiplications and divisions are done element-wise and $\mathcal{F}^{-1}\{.\}$ denotes the inverse STFT operation.

The optimisation problem (4) essentially solves the problem of fitting $\boldsymbol{Y}$ onto $\boldsymbol{D} = [\boldsymbol{D}_1\ \boldsymbol{D}_2]$. Fitting implies any part of the mixed signal spectrogram $\boldsymbol{Y}$ can be represented by either $\boldsymbol{D}_1$ or $\boldsymbol{D}_2$ so long as the total error in reconstruction is minimised. But the task of separation is different. Separation requires the signals to get represented dominantly by its own model. The discriminiative structures are thus to be imposed during training.

2.1. Parameters for quantifying effective separation

Since, an effective discriminative training is the answer to effective separation, a natural question to ask is: how do we parametrically quantify discrimination? To this end, we define a few ratio based parameters that naturally quantify the separation that can be achieved using the models obtained from training. Since we are interested in recovering only one source at a time from a mixture, rest of the constituents are treated as "interferers"[1]. We denote the source and interferer models by $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$ which are $N \times k_s$ and $N \times k_n$ matrices. We denote the spectrograms of training data as $\boldsymbol{S}_s$ and $\boldsymbol{S}_n$ which are $N \times J_s$ and $N \times J_n$ matrices. Here, N is dependent on the FFT size.

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. Solving (4) when $\boldsymbol{Y}$ is the source signal $\boldsymbol{S}_s$.

$$\boldsymbol{C}_{ss}, \boldsymbol{C}_{sn} = \underset{\boldsymbol{C}_{ss},\boldsymbol{C}_{sn}}{\text{argmin}}\ D_{KL}(\boldsymbol{S}_s \| \boldsymbol{D}_s\boldsymbol{C}_{ss} + \boldsymbol{D}_n\boldsymbol{C}_{sn})$$
$$E_{ss} = \| \boldsymbol{D}_s\boldsymbol{C}_{ss} \|_F \gg E_{sn} = \| \boldsymbol{D}_n\boldsymbol{C}_{sn} \|_F \tag{7}$$

   Define the source energy ratio $r_s = \frac{E_{ss}}{E_{sn}}$.

2. Similarly, when $\boldsymbol{Y} = \boldsymbol{S}_n$, on solving (4),

$$\boldsymbol{C}_{ns}, \boldsymbol{C}_{nn} = \underset{\boldsymbol{C}_{ns},\boldsymbol{C}_{nn}}{\text{argmin}}\ D_{KL}(\boldsymbol{S}_n \| \boldsymbol{D}_s\boldsymbol{C}_{ns} + \boldsymbol{D}_n\boldsymbol{C}_{nn})$$
$$E_{nn} = \| \boldsymbol{D}_n\boldsymbol{C}_{nn} \|_F \gg E_{ns} = \| \boldsymbol{D}_s\boldsymbol{C}_{ns} \|_F \tag{8}$$

   Define the interferer energy ratio $r_n = \frac{E_{nn}}{E_{ns}}$.

Clearly, a high value of the source energy ratio r_s would ensure better reconstruction of the source and a 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,

[1] We adopt the to-each-its-own strategy for all the constituents

to reduce interference in the recovered source. We define an error ratio r_e

$$r_e = \frac{\frac{1}{J_n}\sum_{j=1}^{J_n} \|\boldsymbol{s}_n^j - \boldsymbol{D}_s \boldsymbol{c}_{ns}^j\|_2}{\frac{1}{J_s}\sum_{j=1}^{J_s} \|\boldsymbol{s}_s^j - \boldsymbol{D}_s \boldsymbol{c}_{ss}^j\|_2} \tag{9}$$

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 vector 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 $\boldsymbol{D}_s$.

Having described the above ratios, we now utilise them to train better discriminative models by improving upon the discriminative NMF framework in [7]. Specifically, we use the ratio parameters described above to obtain $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$ such that a certain level of reconstruction accuracy is provided while ensuring that the interference is restricted. Note that our dimension search framework is general and can be applied to other source models like dictionary, subspace etc as well. 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. However, our algorithm potentially only needs few iterations of the discriminative NMF training.

3. Discriminative NMF algorithm

The discriminative NMF algorithm in [7] trains the models of the two underlying sources jointly, penalising the coherence between the models. We rewrite the optimisation problem in [7]

$$\begin{aligned}\boldsymbol{D}_s, \boldsymbol{D}_n = \underset{\boldsymbol{D}_s, \boldsymbol{D}_n}{\text{argmin}} \quad & D_{KL}(\boldsymbol{S}_s \| \boldsymbol{D}_s \boldsymbol{C}_{ss}) + \\ & D_{KL}(\boldsymbol{S}_n \| \boldsymbol{D}_n \boldsymbol{C}_{nn}) + \lambda \sum_{i,j} (\boldsymbol{D}_s^T \boldsymbol{D}_n)_{ij}\end{aligned} \tag{10}$$

We propose to add the dimension parameters k_s and k_n for the models $\boldsymbol{D}_s$ 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. In a way these ratios act as constraints over and above the optimisation problem (10). Unlike [7], we do not train these models jointly. Instead, we break the optimisation problem into two parts. We first train the source NMF dictionary $\boldsymbol{D}_s$ such that error ratio r_e satisfies a certain threshold r_{th}. Once the source dictionary $\boldsymbol{D}_s$ is trained, we then follow up with training the interferer dictionary $\boldsymbol{D}_n$. The entire process is repeated after exchanging the roles of the source and interferer to recover the other source.

3.1. Finding source model $\boldsymbol{D}_s$ and k_s

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

$$\boldsymbol{D}_s = \underset{\boldsymbol{D}_s, \boldsymbol{C}_{ss}}{\text{argmin}}\, D_{KL}(\boldsymbol{S}_s \| \boldsymbol{D}_s \boldsymbol{C}_{ss}) \tag{11}$$

(11) is solved using multiplicative updates as described below:

$$\boldsymbol{D}_s \leftarrow \boldsymbol{D}_s \otimes \frac{\frac{\boldsymbol{S}_s}{\boldsymbol{D}_s \boldsymbol{C}_{ss}} \boldsymbol{C}_{ss}^T}{\mathbf{1}_s \boldsymbol{C}_{ss}^T}, \quad \boldsymbol{C_{ss}} \leftarrow \boldsymbol{C_{ss}} \otimes \frac{\boldsymbol{D}_s^T \frac{\boldsymbol{S}_s}{\boldsymbol{D}_s \boldsymbol{C}_{ss}}}{\boldsymbol{D}_s^T \mathbf{1}_s} \tag{12}$$

Here, all the multiplication and division operations are done element-wise. The columns of $\boldsymbol{D}_s$ are normalised after each iteration. $\mathbf{1}_s$ is a matrix of ones with size $N \times J_s$. This search for the dimension k_s of the source dictionary is described in Algorithm 1. Searching for an appropriate k_s is essentially finding that value of k_s such that the error ratio r_e in (9) gets as close as possible to a predetermined threshold r_{th}. Through experiments we have observed that the ratio r_e is generally monotonically increasing with the dimension variable k_s. Hence a binary search over k_s can be performed. For each value of k_s during the search, we solve (11) 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}.

Algorithm 1 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
Solve (11) to train $\boldsymbol{D}_s$ with $k_s$ columns
Output: $\boldsymbol{D}_s$
```

3.2. Finding interferer model $\boldsymbol{D}_n$ and k_n

With the given source NMF dictionary $\boldsymbol{D}_s$, we now determine the interferer $\boldsymbol{D}_n$ and its dimension k_n using ratios r_s and r_n. For a given value of the dimension k_n, $\boldsymbol{D}_n$ is obtained by

$$\boldsymbol{D}_n = \underset{\boldsymbol{D}_n, \boldsymbol{C}_{nn}}{\text{argmin}}\, D_{KL}(\boldsymbol{S}_n \| \boldsymbol{D}_n \boldsymbol{C}_{nn}) + \lambda \sum_{i,j} (\boldsymbol{D}_s^T \boldsymbol{D}_n)_{ij} \tag{13}$$

Here, λ is the regularisation parameter. (13) is solved using multiplicative updates as described below:

$$\boldsymbol{D_n} \leftarrow \boldsymbol{D_n} \otimes \frac{\frac{\boldsymbol{S}_n}{\boldsymbol{D}_n \boldsymbol{C}_{nn}} \boldsymbol{C}_{nn}^T}{\mathbf{1}_n \boldsymbol{C}_{nn}^T + \lambda \boldsymbol{D}_s \bar{\mathbf{1}}_n} \tag{14}$$

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 $\boldsymbol{C}_{ss}$. A search for an appropriate dimension of the interferer model $\boldsymbol{D}_n$ is 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 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. Once the models are trained, equations (4)-(6) (replacing $\boldsymbol{D}_1$ with $\boldsymbol{D}_s$ and $\boldsymbol{D}_2$ with $\boldsymbol{D}_n$) are used to find the estimate of the source signal $\hat{s}_s(t)$.

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

```
Input: $D_s$, $S_n$
$k_n$ = $k_{n,min}$; $in = 1$; {Indicator for search}
while $in == 1$ do
    Solve (13) to find $D_n$ 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: $D_n$
```

4. Results and discussion

4.1. Data

The algorithm was tested for separation of two speech signals and speech-music signals. For the speech case, the algorithm was evaluated on 4 male and 4 female speakers taken from the TIMIT 16k database [13] 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 [14]. 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. Parameters

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. The values of $k_{s,min}$ and $k_{n,min}$ are fixed to be 20 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. Increasing the value of $r_{s,min}$ will give a better reconstruction of the source but will increase the interference in it. λ is fixed at 100.

4.3. Performance of our algorithm

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. We applied our framework on NMF dictionaries as in [7], here referred to as RNMF which solves (10). We term our method of application of the discriminative framework on NMF dictionaries as DF-NMF. Results show that the quality of separation can be improved with the framework. We also compare our results with the dictionary learning based approach in [12], called SDDL, which extracts the sources in a number of levels to achieve better separation. We compared the performances of the algorithms using SDR and SIR as the evaluation metrics. The BSS evaluation toolbox [15] was used for calculation of the evaluation metrics. Figure 1 show the comparison of SDR for a few speech-speech separation trials while figure 2 show the comparison for a few speech-music cases. 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. Tables 1 and 2 compare the average performance of the algorithms in speech separation and speech-music separation respectively.

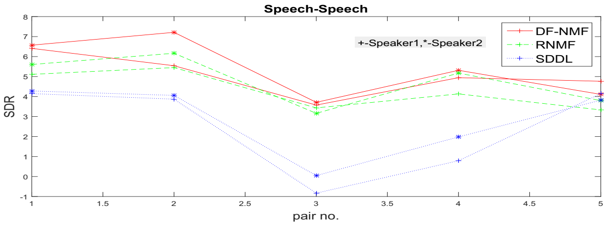

Figure 1: *Comparison for speech-speech case*

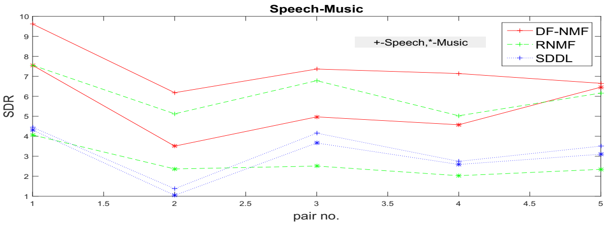

Figure 2: *Comparison for speech-music case*

		DF-NMF	RNMF[7]	SDDL[12]
F+M	SDR SIR	6.46 8.57	5.6 7.01	5.82 8.49
F+F	SDR SIR	4.52 6.45	3.78 4.95	2.38 5.31
M+M	SDR SIR	3.95 5.82	3.56 4.65	2.17 4.68

Table 1: *Average performance for speech-speech separation*

		DF-NMF	RNMF[7]	SDDL[12]
Speech	SDR SIR	7.32 10.23	6.2 13.88	3.06 6.05
Music	SDR SIR	5.04 6.66	2.78 3.2	2.85 5.53

Table 2: *Average performance for speech-music separation*

5. Conclusions

We present a new discriminative framework for the SCSS problem where we suggest an individual optimisation problem for every component of the mixture, while simultaneously searching for an appropriate dimension for the source models as part of their training. We applied our framework on the NMF dictionaries and through experiments show that our method improves the separation performance compared to other existing approaches. Our approach also opens up the possibility of finding theoretical guarantees in source separation.

6. References

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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/.