Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Department of Electronic and Electrical Engineering

Analysis of Robust Principal Component Analysis for Singing Voice Separation

Daniel Woodward
15324849

April 12, 2019

A Final Year Project submitted in partial fulfilment
of the requirements for the degree of
BAI (Computer and Electronic Engineering)

Declaration

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: _____ Date: _____

Abstract

This report is compiled with the intention of documenting an investigation into Robust Principal Component Analysis for Singing Voice Separation across genres with the intention of dynamically selecting parameters to improve it's effectiveness as measured by the normalised signal to distortion ratio.

Originally proposed by Huang et al(1) in 2012, RPCA for SVS uses machine learning to separted a mixed signal into it's singing and accompaniment components. Trying to improve this method based on genre proved to be tough and thus the report is inconclusive on whether it can be improved based on genre alone. It appears that other factors impact the dynamic selection of parameters more than the genre.

Other investigations in this report include frequency analysis for RPCA which yielded a conclusion that higher frequencies tend to be separated more effectively by RPCA.

The vocal training of singers on the datasets used in singing voice separation is also put under the spotlight in this report and evidence is found that in fact that vocal training has a insignificant effect on the effectiveness of RPCA for SVS.

Code used in this project can be found on my GitHub at: https://github.com/Daniel-Woodward

Acknowledgements

I would like to acknowledge and thank my supervisor Naomi Harte for her time, effort and support during the process of my final year project. She is truly a stellar person and supervisor.

I can't give enough thanks to my sister Meghan Woodward and Eimear Kiely for their contributions to the singing voice datasets created during the duration of the project. They allowed for the opportunity to extend the scope of the investigation in RPCA and for that I cannot thank them enough.

Finally thank you to my parents for support me through four years and through this project. Their suppor has been fundemental in seeing me through this project and my time at Trinity College Dublin.

I would like to extend my sincere thanks to all, without the above I would never have finished this project.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem Statement and Aims

People have an inate ability to disguinish between the different musical signals that combine to form a song. Intuitively able to separate the human voice from a backing guitar for example. Musicians are capable of even listening to a piece of music and recreating it by ear without any original information on the musical theory of the song. The interpretation of which frequencies belong to what component comes naturally but for computers, the task is significantly more complicated.

The focus of this report is the improvement of singing voice separation from music using robust principal component analysis as proposed by Po-Sen Huang et al(1). The singing voice itself contains information such as the singers vocal range for that song, their sex, the identity of the singer or alternatively the musical accompaniment with the singing voice removed can be used for karaoke. For any automatic process above, effective singing voice separation must be achieved. The current state of the art involves the use of convolutional neural nets which are a supervised method, while the method suggested by Po-Sen Huang et al is a unsupervised method. In this report the aim is to improve RPCA for SVS by using meta information of the music, the genre. Given the genre of a song it may be possible to automate the parameters used in RPCA for SVS. Specifically the K value used in RPCA partly for trading the amount of information between the recovered singing voice and the musical accompaniment.

RPCA for SVS is a machine learning technique that is suitable for separation of monaural songs. It takes advantage of the bigger variance in the singing voice signal as compared to the musical accompaniment signal with a repeating underlying song structure. Robust Principal Component analysis is a matrix factorisation algorithm to seperate a matrix into a spare matrix (singing) and an underlying low rank matrix (accompaniment)(1). Using genre information on the songs, the improvement of the singing voice separation can by parameter automation will hopefully increase the effectiveness of RPCA for this application.

1.2 Report Limitations

1.2.1 Datasets

The primary datasets using in singing voice separation are the MIR-1K dataset and the DSD100 dataset. In this report and the papers discussing singing voice separation these datasets are representitive of the songs that RPCA should be optimised for. However, there are characteristics of these datasets which differentiate them from an ideal dataset.

Table 1.1: Dataset Charateristics

Name	Genre	No. of Songs	Monaural	Notes	Sample rate (samples/second)
DSD100	Mixed	102	No	Stereo Recording	44100
MIR-1K	Pop	100	Yes	Low sample rate Poor Quality Recordings	16000

Table 1.1 shows some of the characteristics of the two Datasets. Consider the applications for singing voice separation and the songs to which it would be applied. These songs would be typically mainstream that a created and curated by professionals. The MIR-1K dataset seems to be ubiquitous among the papers that using RPCA for SVS. However, the MIR-1K dataset shows some significant differences from this typical test song. Firstly, the MIR-1K dataset uses a low sample rate of 16000 samples per second. The standard sample rate used is 44100 per second as the Nyquist theorem shows that the sample rate needed for a continuous signal to store information on frequencies of N Hz is approximately $2.2 \times (N)$. 44100 samples per second is used as the accepted range of human hearing is from 20Hz to 20000Hz. Hence, 44100 is sufficient for recording human audiable frequencies. A sample rate of 16000 may therefore be inadaquate as a sample rate for a dataset meant to represent the typical real world song that RPCA for SVS would be applied too.

It should also be noted that the authors of the MIR-1K dataset admit "Most of the singers are amateur and do not have professional music training". Again this may be indicative of a dataset that is not comparable to professionally produced music. This is investigated in this project in chapter 4.

RPCA for SVS, as mentioned previously is only appropriate on monaural data. Monaural music sounds as if it is produced from one position. It is a single signal as opposed to a stereo music signal which contains two signals. The DSD100 dataset represent a more professional and typical set of songs but at the expense of being stereo. To overcome this the music is linearly combined in order to form a monaural song.

1.2.2 Genre Subjectivity

The main focus of this report is to investigate RPCA across different genres. While genres have different characteristics that can be quantified the overall genre of a song is somewhat open to interpretation.

For example in 2019, Old Town Road by lil Nas X was removed from the Billboard hot country chart as it "does not embrace enough elements of today's country music to chart in its current version"[1]. This is contradicted by the accompaniment of the song heavily comprising of a sampled country banjo musical phrase and the lyrics comprising of themes found in country music. There are hip-hop characteristics too, like the use of 808 drums, 16th hit hats and the rapping delivery of lyrics. Thus, there is an argument that the song belongs to either or both genres depending on subjective interpretation.

Further on, when parameters are selected based on genre this causes a problem as if a song straddles multiple genres how should it be decided which parameter values to use? The genres in the DSD100 dataset are labelled and the genre information does show multiple genres on some tracks. Track 10, A place for us by Carlos Gonzalez is labelled as Indie Pop/Rock. For the purposes of experiments described in this report Pop is chosen as the genre when applicable. This is because the popular music genre can be defined as being an amalgamation of several genres into whatever the popular genre is at a given time. I.e. pop music may drawn on characteristics of several genres anyway.

Finally, each genre itself can have several sub genres. These sub genres usually exhibit characteristics of their parent genre and then next descriptor offers an even more in depth description. For the purposes of this report only the main genre is considered and songs are catagorised as such. E.g. Indie Pop is considered Pop music.

1.2.3 Number of songs per genre

Ideally there would be datasets as similar as possible but containing songs of different genres. While there is enough data to test pop music as the MIR-1K dataset comprises totally pop music the analysis for rock genre can only be based on a limited number of songs from the DSD100 dataset.

[1] https://www.cbsnews.com/news/old-town-road-lil-nas-x-billboard-removes-song-from-hot-country-critics-question-race-factor/

2 Technical Background

2.1 Music Signals and Theory

Audio is perceived when the pressure of air changes this causes the eardrum to also move at that same frequency. This induces a small electrical signal in the ear that travels to the brain where it is interpreted as sound. Speakers work by vibrating backwards and forwards at these frequencies in order to make the excite the molecules in the air vibrate at the require frequencies. This is how all audio signals including music are transferred from speakers to the brain. Music can be assessed in the same way any audio signal can. In the vast majority of cases and for all cases in computers they are stored as a set of digital values that are converted to analogue signals before they are sent to the speakers. The generally accepted range of human hearing is 20-20kHz. Therefore, these are the range of values that a audio signal should be able to store and then resynthesis from a sampled signal.

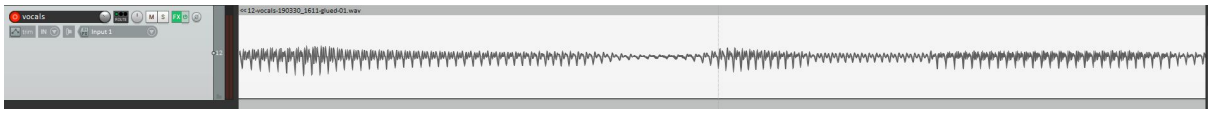

Figure 2.1: Example of an audio signal

The above figure shows an example of an vocal audio signal. The signal however is not sinusoidal or match any obvious shape, this is due to the overtones in the singing voice. A singing voice is typically said to be at a certain frequency, E.g. at 220Hz. The human voice not only produces this fundamental frequency but also a number of overtones. These overtones are integer multiples of the fundamental frequency. The ratio of the amplitude of these overtones give qualities to a voice (the timbre) and are also how different instruments produce different sounds I.e why a piano sounds different to a guitar when the same note is played.

When musician composes a musical song the different instruments are combined in order to produce the final mixture. Now even if the fundamental frequencies of all the individual parts are known it is necessary to classify the overtones in order to recover the human voice.

2.2 RPCA for singing voice separation

2.2.1 Overview

Robust Principal Component Analysis for singing voice separation separates a matrix representing a spectrogram of the input signal into a low rank and sparse matrix. Here an assumption is made that the variance in the singing voice is higher than the variance in the accompaniment. The underlying intuition is that the repetitiveness of music allows the accompaniment to occupy a low rank subspace. Hence when the input spectrogram matrix is separated, the idea is that the singing voice is classified into the sparse matrix and the accompaniment into the low rank matrix. (1)

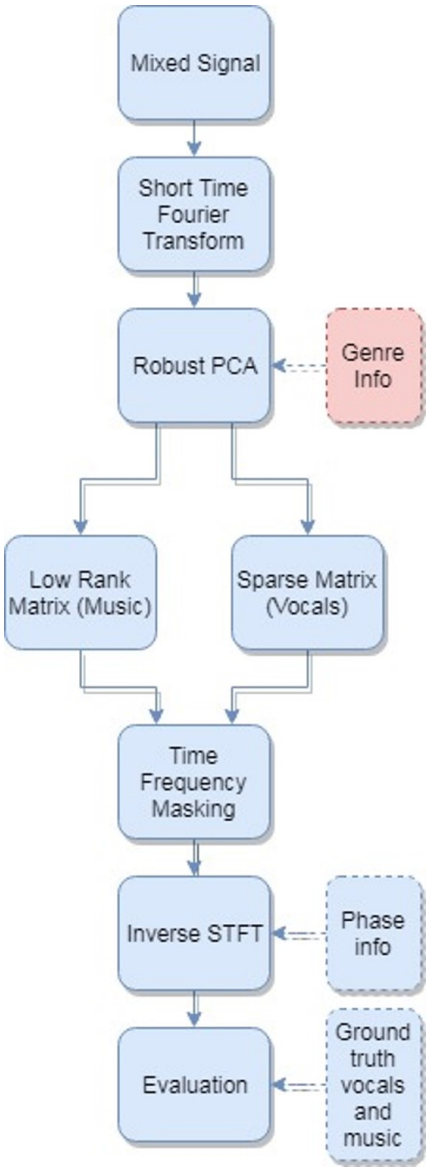

Figure 2.2: Top View of RPCA for SVS

2.2.2 Mixed signal and Short time Fourier transform

The input for the RPCA method uses the mixed song signal sampled at 44100 samples per second which contains whole song to be separated. Firstly a short time Fourier transform is performed on the input signal. The short time Fourier transform produces the sinusoidal and phase content of a signal over time. The output is a matrix where each column represents one N-point fft. Then the window is offset by $\frac{\texttt{window size}}{2}$ in this case and this is repeated for the input signal length. The output is a matrix with the x axis representing the time and the y axis representing the frequency.

In Image 2.3 the input signal shown as a spectrogram. The horizontal red lines are indicators that at that frequency and time there exists an audible signal. Even from visual analysis of these diagrams it is clear that there is some source in the song clip increasing in frequency up to around 8 seconds before decreasing in frequency again. Note also the overtones of this rising and falling signal are visible but they are less prominent as the fundamental frequency (F0) is the dominant frequency.

RPCA intuitively tries to separate the input spectrogram into two matrices with the low rank matrices having as few columns and rows as possible with values while maintaining a sparse matrix that contains as few values as possible. Hence, the low rank accompaniment spectrogram appears with many perceived horizontal and vertical lines observed in red. These are the frequencies at given times that remain more constand I.e. have less variation. Consider now the sparse matrix, there are relatively few entries denoted by the vast amount of blue in the spectrogram representing frequencies at times with no amplitude. The values classified into the sparse matrix should correspond to the input source signal components with greater variation, our assumption for the vocal frequencies.

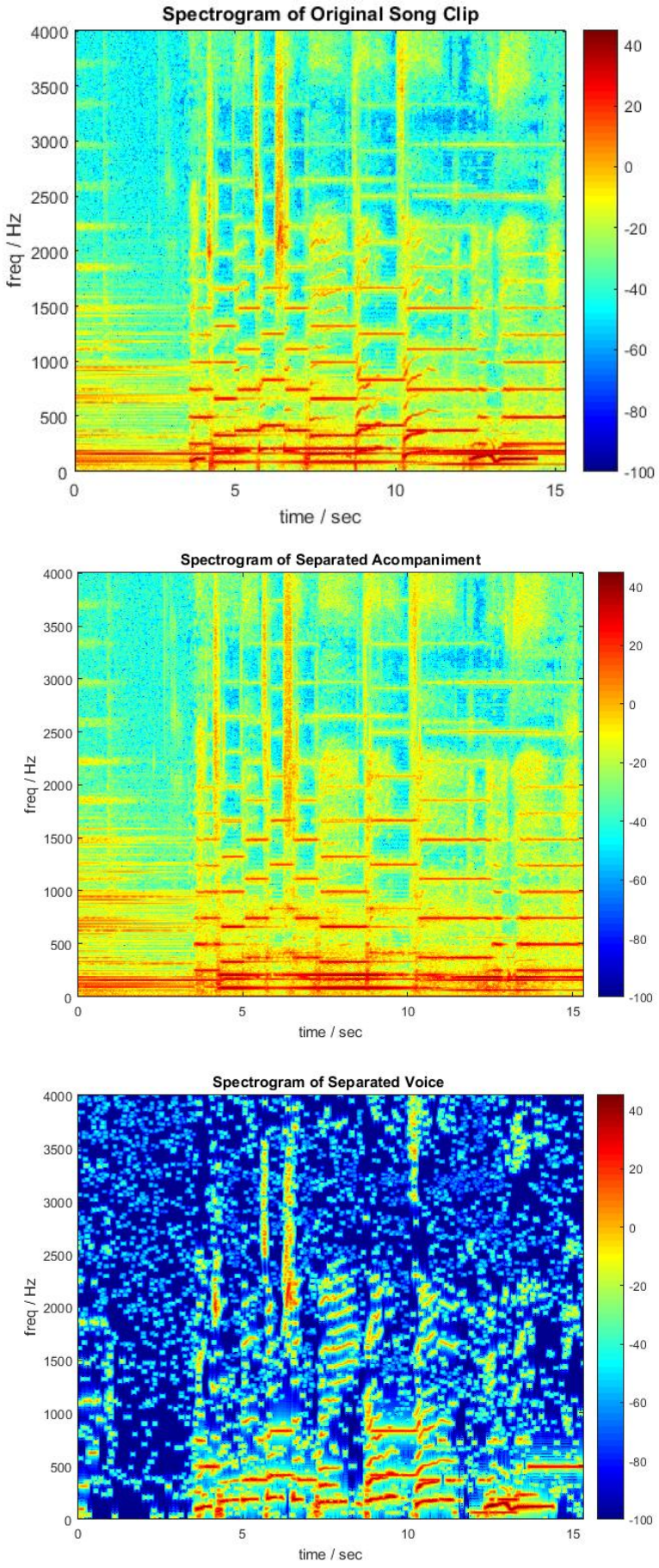

Figure 2.3: Spectrograms before and after RPCA separation

2.2.3 Robust Principal Component Analysis

The input matrix is separated in this step into the sparse and low rank matrices. It was originally proposed by Candes et al as a way of reconstructing low rank matrices with corrupted entries. This approach is called Principal component persuit and tries to solve the convex optimisation problem:

$$\begin{aligned} &\texttt{minimize } \|L\|_* = \lambda\|S\|_1 \\ &\texttt{subject to } L + S = M \end{aligned} \tag{1}$$

Here L is the low rank matrix used to store the frequency and time information corresponding to the accompaniment and similarly S represents the sparse matrix information corresponding to the singing voice and M is the mixed source such that $M \in R^{n1 \times n2}$, $S \in R^{n1 \times n2}$, $L \in R^{n1 \times n2}$ where n1 and n2 are the size of the columns and rows of the input spectrogram matrix respectively. $\|.|_*$ and $\|.|_1$ denote the nuclear norm and the L1 norm of a matrix.

2.2.4 λ and the K value

The λ value is a trade-off parameter that trades the sparsity of S with the low rankness of L. In effect, this trades the amount of data stored in the S and L matrices with an increase in λ increasing the sparsity of S. Typically in literature this value is taken as:

$$\lambda = \frac{1}{\sqrt{max(n_1, n_2)}} \tag{2}$$

It is suggested that this augmented with a value k that can be changed to fine tune the λ value such that:

$$\lambda = \frac{k}{\sqrt{max(n_1, n_2)}} \tag{3}$$

This K value is the main aspect of RPCA for SVS examined in this report. Recall the main aim is to automate this parameter such that the best signal to distortion ratio is observed(2).

2.3 Time Frequency Masking

After the separation of the input matrix into a low rank and sparse matrix a binary time frequency mask is applied to improve results. The mask M_b is defined as:

$$M_b = \begin{cases} 1, \mid S(m,n) \mid > \mid L(m,n) \mid \\ 0, \quad \texttt{otherwise} \end{cases} \tag{4}$$

for all m = 1...n_1, n = 1...n_2

The computed mask is applied to the spectrogram of the original input which gives the final spectrogram from RPCA for SVS such that:

$$\begin{aligned} M_{singing} &= M_b(m,n)M(m,n) \\ M_{accompaniment} &= (1 - M_b(m,n))M(m,n) \end{aligned} \tag{5}$$

Finally the inverse short time fouier transform is applied to our newly calculated spectrograms to resynthesis the singing voice and accompnaiment. (1)

2.4 Evaluation of results

Evaluation of results is done using the normalised signal to distortion ratio (NSDR). Consider that a signal can be decomposed into its constituent parts such that:

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t) \tag{6}$$

s_target(t) = The allow source deformation (The singing voice)
e_noise(t) = Perturbing noise not from the source (n/a in this case)
e_interf(t) = unwanted interference from source (Incorrectly classified matrix values)
e_artif(t) = artifacts introduced from the separation algorithm

The Signal to distortion ratio is then described as:

$$SDR =: 10log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}(t) + e_{noise}(t) + e_{artif}(t)\|^2} \tag{7}$$

However this metric is heavily biased by the amplitude of the singing voice compared to amplitude of the overall accompaniment. Instead a metric that measures the improvement of the SDR would be more appropriate. Hence, the normalised SDR is used. This is computed by:

$$NSDR =: SDR_{resynthesised} - SDR_{source} \tag{8}$$

Now the original resynthesised voice is compared to the original sources using the SDR. This gives a better estimate of the effectiveness of the singing voice separation technique for that song. The BSS eval toolkit for Matlab contains the tools and functions necessary for these calculations(3).

2.5 Current RPCA for Singing Voice Separation

Yukara Ikemiya et al proposed an addition to RPCA for singing voice separation using the addition of a F0 (fundemental frequency) vocal estimation to make a harmonic mask. The Harmonic mask is used in conjunction with the RPCA mask to produce the intergrated mask for singing voice separation(4).

The F0 estimation is computed applying the Viterbi search on F0 saliency spectrograms which produces the most likely sequence of hidden states. The sequence in this case is the F0 frequency with time.

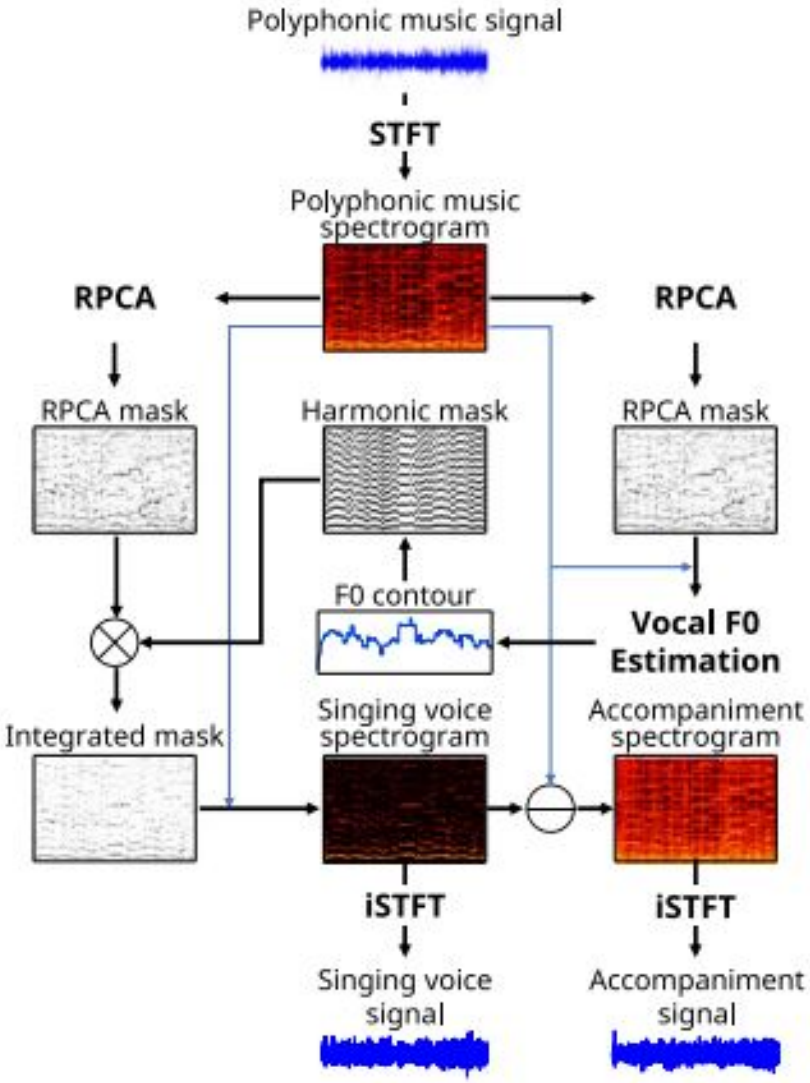

Figure 2.4: Top View of RPCA for SVS with F0 estimation editing

The figure above shows the top level design of the method which involved the following steps:

- Short time Fourier Transform to produce a spectrogram
- RPCA to make a separation mask
- vocal F0 contour estimated by applying the Viterbi search on F0 saliency spectrogram to make harmonic mask
- Both masks integrated via element wise multiplication to obtain final integrated mask. . .

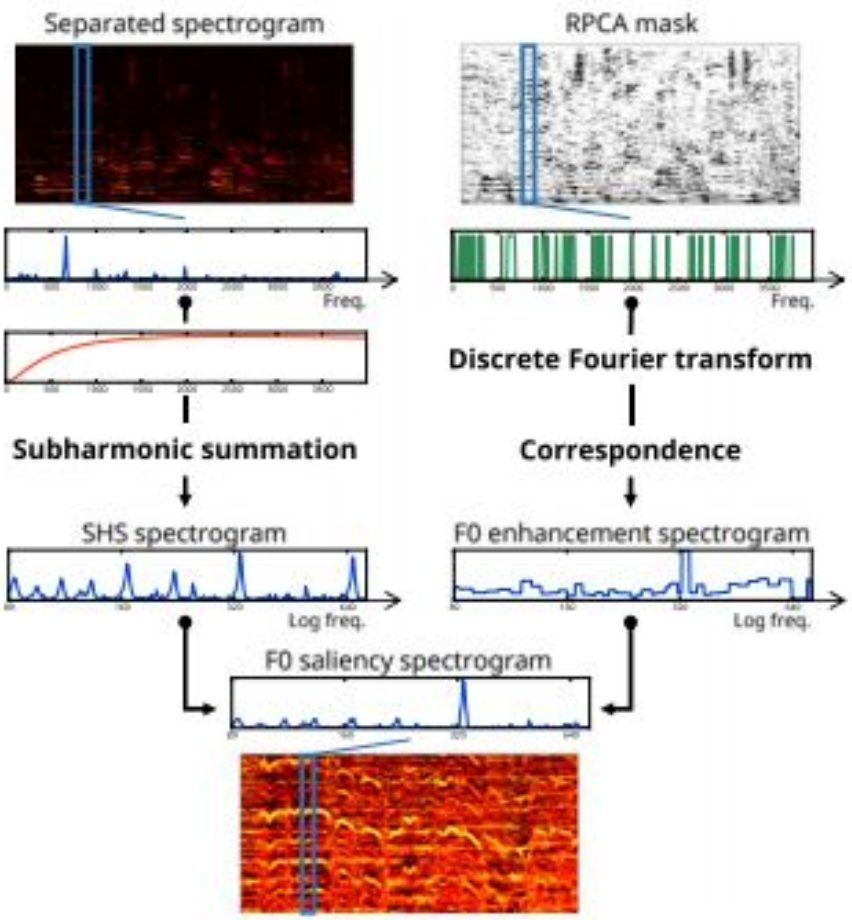

Figure 2.5: F0 estimation algorithm

Figure above shows the F0 saliency spectrogram calculation. The separated spectrogram comes from the RPCA mask applied to the original mixed signal matrix. Subharmonic summation is then used to produce the SHS spectrogram.

This results in the improvement of RPCA as

MIR-1K Dataset	
GNSDR	5.79

2.6 State of the art

The current state of the art in singing voice separation is proposed by Naoya Takahashi et al in conjunction with Sony cooperation in "AN EFFICIENT COMBINATION OF CONVOLUTIONAL AND RECURRENT NEURAL NETWORKS FOR AUDIO SOURCE SEPARATION"(5).

In the paper the architecture is a combination as suggested in two previous papers. Mixtures of MMDenseNet with LSTM are tested in three configurations and compared to several previously used architectures with sequential LSTM after dense layer model providing the best results.

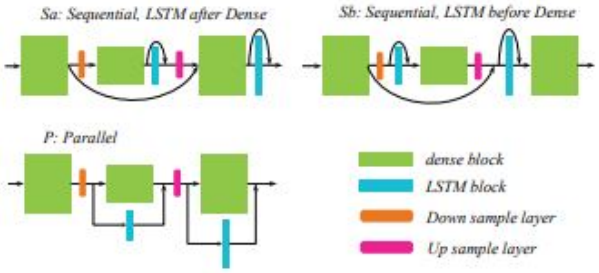

Figure 2.6: New Architectures

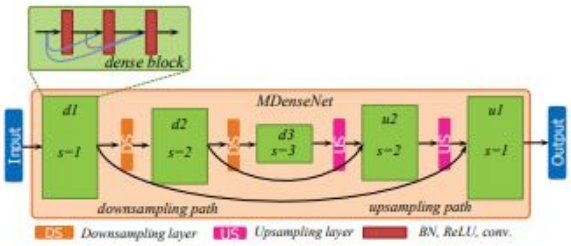

Figure 2.7: MDenseNet Archtitecture

MDenseNet as shown in figure 2.4 is the base architecture used for the MMDenseNetLSTM. The input is split into several bands, in the paper cut offs of 4.1kHz and 11kHz with an addition input of the full band. The output from each band is then concatenated and a final dense block integrates the features from these bands to produce the final output.

The improvement in state of the art comes from the introduction of the LSTM blocks. The three different architectures are tested and the results show Sa: sequential after LSTM produces the best results.

SDR in Db				
Bass	Drums	Other	Vocals	Accompaniment
3.73	5.46	4.33	6.31	12.73

3 Dynamic Parameter Selection for RPCA

3.1 Parameter selection overview

There are several parameters associated with RPCA. For example, the short time fourier transform has several such as window length. For this report they take as constants however as shown in table 3.1. The parameter dynamically chosen is the value of K though which recall is used to trade the low rankness of the low rank matrix and the sparsity of the sparse matrix. Typically, K is taken at a value of K=1. However this report investigates the ideal value of this K value for difference genres and the impact that an ideal value of K will have on the normalised signal distortion ratio for each song and then across each genre.

Table 3.1: Parameters for STFT

Parameter	Value
Window Size	1024
No. fft points	1024
Hop size	512

The DSD-100 dataset contains 100 songs. for each of these pop songs a 'K test' was performed where the NSDR was calculated for the resynthesised voice. For most songs this produces a graph that shows a significant exponential increase in NSDR before a sharp fall in the results as seen in figure 3.1 below.

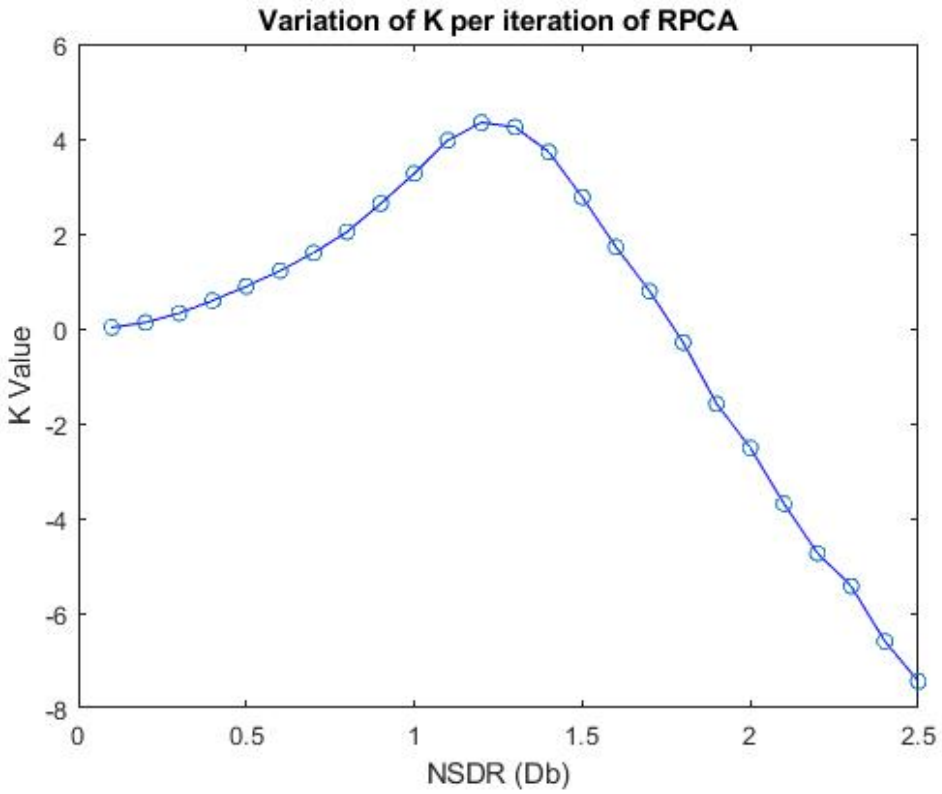

Figure 3.1: Typical shape of NSDR with varying K value

3.1.1 Implementation of RPCA

For this investigation the code provided by Po Sen Huang is used for the implementation of RPCA for SVS [1]. This code provides most of the functionality needed for this project. The implementation of RPCA is done using inexact augmented lagrangian multipliers. This is an approximation of RPCA as it significantly reduces the computation power needed.

3.2 DSD-100 Pop

This K test is then performed on the 26 pop songs available in the DSD-100 dataset to find on average the best K value which is then used to recalculate the NSDR for the improved K value.

[1] availible at https://github.com/posenhuang/singingvoiceseparationrpca

Name	Duration	NSDR (dB)	K Value
003 Actions - One Minute Smile	2'44"	0.877169682	1
004 Al James - Schoolboy Facination	3'21"	0.670039068	0.9
006 Atlantis Bound - It Was My Fault For Waiting	4'28"	3.389242451	1.7
009 Bobby Nobody - Stitch Up	3'41"	3.257503893	1.2
010 Carlos Gonzalez - A Place For Us	4'10"	0.52417	0.3
013 Drumtracks - Ghost Bitch	5'57"	0.229494421	0.6
019 James Elder & Mark M Thompson - The English Actor	3'25"	3.35153457	1.5
024 Leaf - Come Around	4'24"	6.618526133	0.9
025 Leaf - Wicked	3'11"	3.929908605	2
032 Raft Monk - Tiring	3'32"	0.298574041	1
037 Speak Softly - Broken Man	4'07"	-0.001923744	0.1
041 The Mountaineering Club - Mallory	4'04"	3.052122547	1.6
042 The Wrong'Uns - Rothko	3'22"	2.57048661	1.1
046 Triviul - Dorothy	3'01"	-0.002126491	0.1
051 AM Contra - Heart Peripheral	3'30"	5.018810398	1.4
053 Actions - Devil's Words	3'17"	1.703838756	1.1
054 Actions - South Of The Water	3'11"	4.826183219	1.3
058 Ben Carrigan - We'll Talk About It All Tonight	4'15"	0.01683526	1.1
072 Jay Menon - Through My Eyes	4'13"	0.960919612	1.1
075 Leaf - Summerghost	3'52"	2.648496004	0.9
078 Moosmusic - Big Dummy Shake	3'19"	4.078291983	1.7
083 Remember December - C U Next Time	3'46"	1.977439094	1.2
096 Triviul - Angelsaint	3'51"	1.273486747	1.1
098 Triviul feat. The Fiend - Widow	3'43"	3.897034751	1.2

Table 3.2: Pop music ideal K values and NSDR (dB)

To select the best value of NSDR the best values for each song are taken and then a mean average is formed.

Mean average = 1.0875

Name	NSDR K=1 (dB)	NSDR K=1.1 (dB)	Difference (dB)
003 Actions - One Minute Smile	0.877169682	0.866233405	-0.010936277
004 Al James - Schoolboy Facination	0.549611576	0.253238214	-0.296373362
006 Atlantis Bound - It Was My Fault For Waiting	1.527202658	1.69504241	0.167839752
009 Bobby Nobody - Stitch Up	2.626494282	3.082630242	0.45613596
010 Carlos Gonzalez - A Place For Us	0.310715819	0.22498257	-0.085733249
013 Drumtracks - Ghost Bitch	-0.068355843	-0.378704939	-0.310349096
019 James Elder & Mark M Thompson - The English Actor	2.733443915	2.826291009	0.092847094
024 Leaf - Come Around	6.537474007	6.097020496	-0.440453511
025 Leaf - Wicked	2.669375902	2.753282522	0.08390662
032 Raft Monk - Tiring	0.298574041	0.111363371	-0.18721067
037 Speak Softly - Broken Man	-3.102466113	-4.424797973	-1.32233186
041 The Mountaineering Club - Mallory	0.848303769	1.332364159	0.48406039
042 The Wrong'Uns - Rothko	2.42876546	2.57048661	0.14172115
046 Triviul - Dorothy	-0.915717893	-1.294256167	-0.378538274
051 AM Contra - Heart Peripheral	1.991067343	2.807615515	0.816548172
053 Actions - Devil's Words	1.645477388	1.703838756	0.058361368
054 Actions - South Of The Water	3.083418453	3.921844703	0.83842625
058 Ben Carrigan - We'll Talk About It All Tonight	-0.015038497	0.01683526	0.031873757
072 Jay Menon - Through My Eyes	0.894078723	0.960919612	0.066840889
075 Leaf - Summerghost	2.58438729	2.409816802	-0.174570488
078 Moosmusic - Big Dummy Shake	1.168218586	1.386567828	0.218349242
083 Remember December - C U Next Time	1.896780756	1.973266245	0.076485489
096 Triviul - Angelsaint	1.149426818	1.273486747	0.124059929
098 Triviul feat. The Fiend - Widow	0.008389374	-0.029073573	-0.037462947

Table 3.3: DSD100 Pop comparison table

Average Difference = 0.413496

There is a significant improvement if the average best K value is selected in this case.

Hence for the DSD100 dataset there is marked improvement in NSDR by using an idealised calculated value of K.

3.3 DSD-100 Rock

Similarly there are 44 rock songs in the DSD-100 dataset. A K-test is performed for each for each of the rock songs similarly to the pop songs.

Average Ideal K value = 1.2

Name	Length	NSDR (dB)	K value
007 BKS - Too Much	4'04"	6.187791455	1
008 Bill Chudziak - Children Of No-one	3'51"	1.318032676	0.7
011 Cnoc An Tursa - Bannockburn	4'54"	0.659439351	0.7
012 Dark Ride - Burning Bridges	3'53"	7.740392391	1
014 Fergessen - Back From The Start	2'43"	0.649157537	1.2
015 Fergessen - The Wind	3'12"	2.351736365	0.8
016 Forkupines - Semantics	4'34"	0.017506707	0.3
018 Hollow Ground - Ill Fate	2'22"	7.025457944	0.9
022 Johnny Lokke - Promises & Lies	4'46"	7.901086747	1.2
023 Jokers, Jacks & Kings - Sea Of Leaves	3'11"	2.792081271	1.7
026 Louis Cressy Band - Good Time	4'53"	1.306153457	0.7
028 Motor Tapes - Shore	4'07"	2.681439215	0.8
029 Nerve 9 - Pray For The Rain	5'44"	1.345261648	0.6
031 Phre The Eon - Everybody's Falling Apart	3'44"	2.791227417	1.2
035 Signe Jakobsen - What Have You Done To Me	2'57"	2.228652137	0.7
038 Spike Mullings - Mike's Sulking	4'08"	1.172431394	1.3
039 Swinging Steaks - Lost My Way	4'58"	2.990331953	1.5
040 The Long Wait - Back Home To Blue	4'15"	3.895899023	1.1
043 Timboz - Pony	4'07"	0.041625012	1.6
045 Traffic Experiment - Sirens	6'55"	0.850637681	0.9
047 Voelund - Comfort Lives In Belief	3'18"	0.84222656	1
048 We Fell From The Sky - Not You	3'28"	1.343889738	2.5
049 Young Griffo - Facade	2'48"	3.332281691	1.5
050 Zeno - Signs	3'54"	1.27877844	1
055 Angels In Amplifiers - I'm Alright	3'00"	2.844439026	1.2
057 BKS - Bulldozer	6'47"	-0.069760015	0.1
059 Black Bloc - If You Want Success	6'39"	3.202103594	1
060 Buitraker - Revo X	4'36"	1.41663893	2.3
063 Detsky Sad - Walkie Talkie	3'10"	2.989519207	1.8
065 Fergessen - Nos Palpitants	3'18"	0.951981812	1.2
066 Flags - 54	5'15"	6.617109791	2
069 Hollow Ground - Left Blind	2'39"	5.291264812	1.4
073 Johnny Lokke - Whisper To A Scream	4'16"	6.740389949	1.7
074 Juliet's Rescue - Heartbeats	4'28	7.389802335	2.3
079 Mu - Too Bright	4'32"	-0.005995369	0.1
080 North To Alaska - All The Same	4'08"	4.255942425	1.8
090 The Doppler Shift - Atrophy	5'31"	5.356978934	1.7
091 The Long Wait - Dark Horses	4'54"	1.913184113	1.1
093 Tim Taler - Stalker	3'34"	0.140382364	0.4
094 Titanium - Haunted Age	3'56"	0.530992278	1.1
095 Traffic Experiment - Once More (With Feeling)	6'57"	1.768606346	1.2
098 Wall Of Death - Femme	3'47"	0.421146283	1.8
099 Young Griffo - Blood To Bone	4'03"	3.04586219	1.5
100 Young Griffo - Pennies	4'26"	2.134457812	1.3
067 Georgia Wonder - Siren	7'10"	1.26498912	1.1

Table 3.4: Rock music ideal K values and NSDR (dB)

Name		Length	NSDR K=1 (dB)	NSDR K=1.2 (dB)	Diff
007 BKS - Too Much	Rock	4'04''	6.187791455	5.753718769	-0.434072686
008 Bill Chudziak - Children Of No-one	Rock	3'51''	0.969053339	0.41438359	-0.554669749
011 Cnoc An Tursa - Bannockburn	Rock	4'54''	0.00089277	-1.145264005	-1.146156775
012 Dark Ride - Burning Bridges	Rock	3'53''	7.740392391	7.516603047	-0.223789344
014 Fergessen - Back From The Start	Rock	2'43''	0.563288437	0.649157537	0.085869099
015 Fergessen - The Wind	Rock	3'12''	1.851407	0.758670395	-1.092736604
016 Forkupines - Semantics	Rock	4'34''	-1.550340231	-2.337592449	-0.787252218
018 Hollow Ground - Ill Fate	Rock	2'22''	6.786988988	5.992400156	-0.794588833
022 Johnny Lokke - Promises & Lies	Rock	4'46''	1.818110702	2.539776716	0.721666014
023 Jokers, Jacks & Kings - Sea Of Leaves	Rock	3'11''	1.813074843	1.844392311	0.031317468
026 Louis Cressy Band - Good Time	Rock	4'53''	1.11213354	0.857748261	-0.25438528
028 Motor Tapes - Shore	Rock	4'07''	2.496003605	1.948667749	-0.547335857
029 Nerve 9 - Pray For The Rain	Rock	5'44''	0.079848177	-1.143973181	-1.223821358
031 Phre The Eon - Everybody's Falling Apart	Rock	3'44''	2.456682987	2.791227417	0.33454443
035 Signe Jakobsen - What Have You Done To Me	Rock	2'57''	1.709820691	0.857567017	-0.852253674
038 Spike Mullings - Mike's Sulking	Rock	4'08''	0.851703822	1.152741894	0.301038072
039 Swinging Steaks - Lost My Way	Rock	4'58''	1.326511861	2.316955438	0.990443577
040 The Long Wait - Back Home To Blue	Rock	4'15''	3.867504386	3.687041802	-0.180462585
043 Timboz - Pony	Rock	4'07''	-1.088715047	-0.908097395	0.180617652
045 Traffic Experiment - Sirens	Rock	6'55''	0.826376492	0.495609109	-0.330767383
047 Voelund - Comfort Lives In Belief	Rock	3'18''	0.84222656	0.538764055	-0.303462505
048 We Fell From The Sky - Not You	Rock	3'28''	-0.445932744	-1.078320826	-0.632388082
049 Young Griffo - Facade	Rock	2'48''	2.312735506	2.955868653	0.643133147
050 Zeno - Signs	Rock	3'54''	1.27877844	1.000985353	-0.277793087
055 Angels In Amplifiers - I'm Alright	Rock	3'00''	2.536873574	2.844439026	0.307565452
057 BKS - Bulldozer	Rock	6'47''	-0.960523022	-2.100469497	-1.139946475
059 Black Bloc - If You Want Success	Rock	6'39''	3.202103594	2.923949851	-0.278153743
060 Buitraker - Revo X	Rock	4'36''	-0.665052535	-1.416551345	-0.75149881
063 Detsky Sad - Walkie Talkie	Rock	3'10''	1.351431446	1.807799963	0.456368517
065 Fergessen - Nos Palpitants	Rock	3'18''	0.832356316	0.951981812	0.119625496
066 Flags - 54	Rock	5'15''	1.637791026	3.104580867	1.466789842
069 Hollow Ground - Left Blind	Rock	2'39''	2.682135963	4.432027596	1.749891632
073 Johnny Lokke - Whisper To A Scream	Rock	4'16''	1.881969397	3.346727327	1.46475793
074 Juliet's Rescue - Heartbeats	Rock	4'28	0.288208011	0.098679741	-0.18952827
079 Mu - Too Bright	Rock	4'32''	-1.414942178	-2.721024067	-1.306081889
080 North To Alaska - All The Same	Rock	4'08''	2.433764085	3.102931403	0.669167318
090 The Doppler Shift - Atrophy	Rock	5'31''	1.096908253	2.575542141	1.478633888
091 The Long Wait - Dark Horses	Rock	4'54''	1.793411137	1.855706858	0.062295722
093 Tim Taler - Stalker	Rock	3'34''	-1.084360082	-2.347056036	-1.262695954
094 Titanium - Haunted Age	Rock	3'56''	0.468891011	0.510644808	0.041753797
095 Traffic Experiment - Once More (With Feeling)	Rock	6'57''	1.563147191	1.768606346	0.205459154
098 Wall Of Death - Femme	Rock	3'47''	0.008389374	-0.115712585	-0.12410196
099 Young Griffo - Blood To Bone	Rock	4'03''	1.772342569	2.573493743	0.801151173
067 Georgia Wonder - Siren	Rock	7'10''	1.114496342	0.989785744	-0.124710598

Table 3.5: Rock music DSD100 comparison table

Mean average NSDR difference = -2.70056

3.4 MIR-1K Pop

The Mir-1K data set is made up of 110 pop songs. Similar to the previous two tests a ideal NSDR is calculated and the the difference is measured. This is a chance to compare the results across different datasets as the both the DSD100 and MIR-1K dataset both contain pop music.

Table 3.6: Pop music ideal K values and NSDR in MIR-1K (dB)

Name	NSDR (dB)	K value
Ani_1.wav	4.344098425	1.2
Ani_2.wav	1.820374421	1
Ani_3.wav	1.770586396	1
Ani_4.wav	2.531729137	1.2
Ani_5.wav	2.510124462	1.3
Kenshin_1.wav	3.678036556	1.1
Kenshin_2.wav	4.961511013	1.1
Kenshin_3.wav	3.604654857	1.4
Kenshin_4.wav	3.200066776	1.3
Kenshin_5.wav	3.119141886	1
abjones_1.wav	4.388355445	1.3
abjones_2.wav	5.371559074	1.1
abjones_3.wav	1.161262294	0.8
abjones_4.wav	0.690993395	0.7
abjones_5.wav	0.170651139	0.7
amy_1.wav	2.941004177	1.5
amy_10.wav	4.096150095	1.6
amy_11.wav	3.717586018	1.5
amy_12.wav	3.31542805	1.4
amy_13.wav	8.572098377	1.5
amy_14.wav	4.452040813	1.5
amy_15.wav	2.593270273	1.4
amy_16.wav	4.213721545	1.3
amy_2.wav	3.58133823	1.2
amy_3.wav	4.644278556	1.4
amy_4.wav	4.037097941	1.4
amy_5.wav	4.510463557	1.3
amy_6.wav	4.195752879	1.5
amy_7.wav	1.456660358	1.4

amy_8.wav	5.032387867	1.6
amy_9.wav	5.006089704	1.4
annar_1.wav	2.446371939	1.1
annar_2.wav	3.014296104	1.2
annar_3.wav	2.155844656	1.2
annar_4.wav	2.68656764	1.2
annar_5.wav	3.308448694	1.3
ariel_1.wav	2.352063256	1.2
ariel_2.wav	1.839892903	1.2
ariel_3.wav	3.511656115	1.4
ariel_4.wav	0.985497716	1
ariel_5.wav	2.045182001	1.3
bobon_1.wav	3.042693253	1
bobon_2.wav	3.518808342	1.4
bobon_3.wav	5.918285978	1.2
bobon_4.wav	1.333928253	1.1
bobon_5.wav	4.826586395	1.5
bug_1.wav	2.257975012	1.1
bug_2.wav	2.013174955	1.3
bug_3.wav	1.065942009	1.1
bug_4.wav	0.571517472	0.5
bug_5.wav	1.578581254	1.2
davidson_1.wav	1.639917582	1.1
davidson_2.wav	0.488379927	0.7
davidson_3.wav	2.108295346	0.9
davidson_4.wav	1.096444496	1
davidson_5.wav	0.629456148	0.9
fdps_1.wav	1.89866255	1.2
fdps_2.wav	1.430289986	1
fdps_3.wav	3.400992579	1.1
fdps_4.wav	3.08488253	1.2
fdps_5.wav	1.578293024	1.2
geniusturtle_1.wav	0.043302206	0.3
geniusturtle_2.wav	1.307467728	0.7
geniusturtle_3.wav	0.098542048	0.2
geniusturtle_4.wav	0.198101021	0.3
geniusturtle_5.wav	0.055811075	0.2

geniusturtle_6.wav	0.606365114	0.9
geniusturtle_7.wav	0.281480254	0.3
geniusturtle_8.wav	0.187505771	0.3
heycat_1.wav	3.119741878	1.4
heycat_2.wav	2.878119625	1.3
heycat_3.wav	4.554629651	1.5
heycat_4.wav	2.215602424	1.1
heycat_5.wav	2.141910852	1.3
jmzen_1.wav	2.517217995	1
jmzen_2.wav	1.60588718	1.2
jmzen_3.wav	1.415324037	1.1
jmzen_4.wav	3.571003648	1.3
jmzen_5.wav	2.432120114	1.3
khair_1.wav	0.755473841	0.9
khair_2.wav	0.18629382	0.9
khair_3.wav	1.149316905	0.9
khair_4.wav	1.189833747	1
khair_5.wav	0.471352691	0.8
khair_6.wav	0.372428522	0.9
leon_1.wav	5.725082147	1.3
leon_2.wav	3.920234568	1.2
leon_3.wav	0.889473579	0.9
leon_4.wav	2.820348437	1.3
leon_5.wav	3.862763602	1.4
leon_6.wav	4.137481273	1.4
leon_7.wav	1.773901265	1.3
leon_8.wav	4.657226101	1.5
leon_9.wav	5.460918408	1.5
stool_1.wav	1.40505932	1.2
stool_2.wav	5.269648971	1.4
stool_3.wav	1.728574121	1
stool_4.wav	4.928369621	1.1
stool_5.wav	1.343279582	1.1
tammy_1.wav	5.186802467	1.2
titon_1.wav	4.442811996	1.2
titon_2.wav	10.23477197	1.5
titon_3.wav	5.646976092	1.6

titon_4.wav	2.617123061	1.2
titon_5.wav	2.326512723	1.1
yifen_1.wav	3.994104798	1.5
yifen_2.wav	3.176378222	1.6
yifen_3.wav	4.52551859	1.6
yifen_4.wav	3.165100493	1.1
yifen_5.wav	4.276952081	1.4

Mean Average K value = 1.15

Table 3.7: MIR-1K Pop comparison table

Name	NSDR K=1 (dB)	NSDR K = 1.15 (dB)	Diff
Ani_1.wav	3.271284481	4.197076588	0.925792106
Ani_2.wav	1.820374421	1.64696958	-0.17340484
Ani_3.wav	1.770586396	1.651586725	-0.118999671
Ani_4.wav	1.86426199	2.407156507	0.542894517
Ani_5.wav	1.861958657	2.284372007	0.42241335
Kenshin_1.wav	3.308989893	3.690214679	0.381224786
Kenshin_2.wav	4.668644978	4.937837052	0.269192073
Kenshin_3.wav	1.546279548	2.718962389	1.172682841
Kenshin_4.wav	2.383179467	2.965502014	0.582322547
Kenshin_5.wav	3.119141886	2.959701666	-0.159440221
abjones_1.wav	3.055521476	4.087940675	1.032419199
abjones_2.wav	4.93168894	5.388377793	0.456688853
abjones_3.wav	0.863580142	0.07798858	-0.785591562
abjones_4.wav	0.524673071	0.110131723	-0.414541348
abjones_5.wav	-0.482031529	-1.809238012	-1.327206483
amy_1.wav	1.555092481	2.224333221	0.66924074
amy_10.wav	2.472073461	3.162058295	0.689984835
amy_11.wav	2.729372243	3.148283264	0.418911021
amy_12.wav	2.679294038	2.968606449	0.289312411
amy_13.wav	3.233409544	4.912168888	1.678759343
amy_14.wav	3.358709944	3.802673096	0.443963152
amy_15.wav	1.363551234	2.012710531	0.649159297
amy_16.wav	3.034791875	3.915344802	0.880552927
amy_2.wav	3.393898518	3.551455593	0.157557075
amy_3.wav	2.813864081	3.813485643	0.999621562

amy_4.wav	2.060025185	2.971848292	0.911823106
amy_5.wav	3.22681367	4.311859261	1.085045591
amy_6.wav	2.633260475	3.387984929	0.754724453
amy_7.wav	0.943421186	1.278290591	0.334869405
amy_8.wav	3.371296271	4.14549916	0.774202889
amy_9.wav	3.497002873	4.286144454	0.789141581
annar_1.wav	2.153428724	2.374218444	0.22078972
annar_2.wav	2.137854506	2.833449021	0.695594515
annar_3.wav	1.983913593	2.149956652	0.166043059
annar_4.wav	2.199047247	2.633701449	0.434654202
annar_5.wav	2.019114184	2.841309079	0.822194895
ariel_1.wav	2.126595791	2.353700344	0.227104553
ariel_2.wav	1.508314566	1.783690844	0.275376278
ariel_3.wav	2.586239113	3.179926235	0.593687122
ariel_4.wav	0.985497716	0.73691639	-0.248581325
ariel_5.wav	1.501863929	1.890690193	0.388826264
bobon_1.wav	3.042693253	2.481456944	-0.561236309
bobon_2.wav	2.284205973	2.898107683	0.61390171
bobon_3.wav	4.593146317	5.837332331	1.244186014
bobon_4.wav	1.32908851	1.24699738	-0.08209113
bobon_5.wav	2.089437608	3.101315692	1.011878084
bug_1.wav	2.251352388	2.196586512	-0.054765875
bug_2.wav	1.406195876	1.772819351	0.366623474
bug_3.wav	0.998307889	1.058961522	0.060653633
bug_4.wav	-1.160744121	-2.417774621	-1.257030499
bug_5.wav	1.356909518	1.578267726	0.221358208
davidson_1.wav	1.549409217	1.568169302	0.018760085
davidson_2.wav	-0.06700474	-1.158118817	-1.091114076
davidson_3.wav	1.924643911	0.796465833	-1.128178078
davidson_4.wav	1.096444496	0.865444852	-0.230999645
davidson_5.wav	0.524078283	-0.170389775	-0.694468059
fdps_1.wav	1.636503441	1.913523512	0.277020071
fdps_2.wav	1.430289986	1.292314687	-0.137975299
fdps_3.wav	3.369928423	3.298619041	-0.071309383
fdps_4.wav	2.593579916	3.055467838	0.461887922
fdps_5.wav	1.510069755	1.581315738	0.071245983
geniusturtle_1.wav	-2.928563862	-4.221113556	-1.292549694

geniusturtle_2.wav	0.097898979	-1.641934006	-1.739832985
geniusturtle_3.wav	-0.9199092	-2.194417686	-1.274508487
geniusturtle_4.wav	-0.753346426	-1.630674526	-0.8773281
geniusturtle_5.wav	-1.558956122	-2.896054117	-1.337097995
geniusturtle_6.wav	0.471146177	-0.41744382	-0.888589997
geniusturtle_7.wav	-2.389347924	-3.210550912	-0.821202988
geniusturtle_8.wav	-0.217586463	-0.712213802	-0.494627338
heycat_1.wav	2.354074014	2.909696124	0.55562211
heycat_2.wav	2.203634931	2.674462161	0.47082723
heycat_3.wav	2.100060708	3.091883753	0.991823044
heycat_4.wav	2.074023471	2.132608099	0.058584628
heycat_5.wav	1.86592206	2.156485528	0.290563468
jmzen_1.wav	2.517217995	2.283961958	-0.233256037
jmzen_2.wav	1.331737155	1.581052401	0.249315247
jmzen_3.wav	1.349434628	1.405920328	0.0564857
jmzen_4.wav	2.024837343	2.967601941	0.942764599
jmzen_5.wav	1.905294771	2.279327169	0.374032398
khair_1.wav	0.723702317	0.364548956	-0.359153362
khair_2.wav	0.12634089	-0.201846179	-0.328187069
khair_3.wav	1.059186128	0.377762997	-0.681423131
khair_4.wav	1.189833747	0.855054373	-0.334779375
khair_5.wav	0.153793031	-0.476398658	-0.630191689
khair_6.wav	0.363726792	0.071015306	-0.292711486
leon_1.wav	4.347358333	5.428032525	1.080674191
leon_2.wav	3.523338924	3.879811335	0.356472411
leon_3.wav	0.880369454	0.734890245	-0.14547921
leon_4.wav	2.239307217	2.653668472	0.414361255
leon_5.wav	2.534361446	3.280689441	0.746327995
leon_6.wav	3.179474649	3.749543725	0.570069076
leon_7.wav	1.559772165	1.762826988	0.203054823
leon_8.wav	1.866049231	2.503133912	0.637084681
leon_9.wav	2.866418563	3.918899886	1.052481324
stool_1.wav	1.011098318	1.314111565	0.303013247
stool_2.wav	2.446599553	3.690063012	1.243463458
stool_3.wav	1.728574121	1.563830943	-0.164743178
stool_4.wav	4.552773847	4.923471355	0.370697509
stool_5.wav	1.225038472	1.340003437	0.114964965

tammy_1.wav	4.318631348	5.293100095	0.974468748
titon_1.wav	3.74388383	4.482091453	0.738207623
titon_2.wav	2.774631395	5.998618221	3.223986826
titon_3.wav	2.585241844	3.743888501	1.158646657
titon_4.wav	2.213465016	2.602829912	0.389364896
titon_5.wav	2.213713842	2.253506258	0.039792415
yifen_1.wav	2.167147949	2.981073402	0.813925453
yifen_2.wav	0.980914294	1.614288507	0.633374213
yifen_3.wav	1.193481096	1.892509874	0.699028777
yifen_4.wav	2.926055846	3.155545124	0.229489278
yifen_5.wav	2.59287404	3.678223431	1.08534939

Average improvement = 0.237454611 dB

The difference in ideal K is small it is still possible that there is a correlation between the genre and the ideal value of K. However, there is not enough evidence to outright prove within a reasonable doubt that the is connection between datasets when even within datasets the ideal value of K varies so much. Any discrepancy between the pop results for the two datasets could be due not just to genre but to other differences in the datasets. This is explored in the subsequent chapters.

4 Vocal training analysis for RPCA

4.1 Hypothesis

Consider the Mir-1K dataset in which the vocals are sung untrained vocalists whereas the DSD100 dataset appears to have professional vocalists. A study has shown that "untalented" singers are worse at pitch tracking compared to those that have received vocal lessons. RPCA tries to exploit the higher variance in vocal signals compared to musical signals to separate the two. If untrained vocals have more variation than the trained vocals then it may be easier to separate the untrained vocals from accompaniment compared to the trained vocalists.

4.2 Comparative Work

In order to test if there is a discernible difference between the trained and untrained vocalists two datasets need to be created such that the only difference between the two datasets is the trained and untrained vocalist.This led to the creation of two datasets; the MWV-25 dataset and the EKV-34 dataset. The aim is to produce these datasets such that the only difference is the trained and untrained nature of the two vocalists.

The datasets both consist of two female contributors singing 'oo' sounds as described in the international phonetic alphabet for vowel symbols as "u". The participants were invited to listen to a notes of a certain frequency played on piano (4front piano virtual instrument) and then while listening to those notes sing the note too into a microphone. These vocal clips are manipulated to produce a musical phrase. This musical phrase is arpeggio of the root notes 7th chord, a common progression. For example, the C arpeggio consists of the notes C, E, G, and B. These arpeggios are overlaid on the accompaniment for twenty songs. Both of these datasets were recorded using a Shure sm-57 microphone.

4.2.1 EKV-25

The EKV-25 dataset is the singing voice dataset performed by a vocalist with over 8 years of vocal training and is currently pursuing a music degree with a focus on singing. This dataset was recorded in the Cork Institute of Technology, Cork School of Music in their edit room which is acoustically treated. There are 34 musical notes used for the recording of this dataset.

4.2.2 MWV-25

The MWV-25 dataset is a singing voice dataset is performed by somebody with previous singing training. This dataset was recorded in an apartment which was not acoustic treated. There are 25 samples in this dataset.

4.3 Investigation of vocal training on RPCA

RPCA for singing voice separation is performed on the two datasets. The inputs used are only those that appear in both the EKV-34 and MWV-25 datasets in order to provide a fair test and be able to draw fair conclusions from a fair comparison.

4.4 Results

EKV Mean	1.3239
EKV Standard Deviation	0.6651
MWV Mean	1.0547
MWV Standard Deviation	0.5242

4.4.1 Comparitive Results

DSD100 Accompaniment	LSE for each Acompaniment between reco
001 - ANiMAL - Clinic A	0.143184415444818
002 - ANiMAL - Rockshow	0.133881681381575
003 - Actions - One Minute Smile	0.145522725789404
004 - Al James - Schoolboy Facination	0.143236678538537
005 - Angela Thomas Wade - Milk Cow Blues	0.145181913818471
006 - Atlantis Bound - It Was My Fault For Waiting	0.0568502128557234
007 - BKS - Too Much	1.0876091138188
008 - Bill Chudziak - Children Of No-one	0.0807758284768576
009 - Bobby Nobody - Stitch Up	0.190278014342741
010 - Carlos Gonzalez - A Place For Us	0.0874888310774956
011 - Cnoc An Tursa - Bannockburn	0.309614323855861
012 - Dark Ride - Burning Bridges	0.524681637689903
013 - Drumtracks - Ghost Bitch	0.109618007706322
014 - Fergessen - Back From The Start	0.110519604672443
015 - Fergessen - The Wind	0.125754203921844
016 - Forkupines - Semantics	0.0618772454684776
017 - Girls Under Glass - We Feel Alright	0.170883525285559
018 - Hollow Ground - Ill Fate	0.363469909547578
019 - James Elder _ Mark M Thompson - The English Actor	0.0414894341226261
020 - James May - Dont Let Go	0.137884266152679

Table 4.1: Comparison MWV and EKV recovered vocals

4.4.2 Comparitive Analysis

Table 4.1 shows the least squares error between the EKV and the MWV arpeggios combined with the DSD100 dataset. For 17 out of the 20 DSD accompaniments the least square error for the the EKV and the MWV datasets is less than 0.3. This represents no significant increase in the effectiveness of RPCA in the two datasets.

Paired T-tests were performed on the recovered vocals NSDR using the in built function in Matlab for each DSD100 accompaniment and again in 17 of 20 accompaniments not enough evidence was found to conclude that there is a difference in the effectiveness of RPCA for the two datasets and hence, no difference in the effect of vocal training on the SVS results in most cases.

5 Frequency analysis for RPCA

5.1 overview

From the intial investigate of RPCA several questions arose aside from the main goal of the difference in RPCA across genres. The range of frequencies that a adult male and adult female sing at differs from 85-180Hz and 165-255Hz respectively. This difference in sing frquencies may impact the effectiveness of RPCA for singing voice separation. In this section the aim is to investigate whether the difference in frequencies results in a significant difference in NSDR. In order to do this the MWV-34 dataset is used to create similar input ideally differing only in frequency only.

5.2 RPCA Testing

Similarly to the vocal training section the data is manipulated into arpeggiated segments. The different segments are Major 7 arpeggios which span across 15 different root notes. These are then combined with twenty song accompaniments from the DSD100. Then by using the NSDR as a metric the results should indicate whether there is sufficient evidence to suggest that there is a difference in the results to conclude that sung frequencies make an impact on RPCA for singing voice separation.

5.3 Testing

For each of the song clips the different accompaniments are overlaid with all the singing arpeggios. This yields a table of NSDR v Root note frequency. An example is shown below.

As shown there is a significant slope of 0.0164 which while low suggests that for this graph over the typical range of a female vocalist, an estimated 90Hz which would result in a

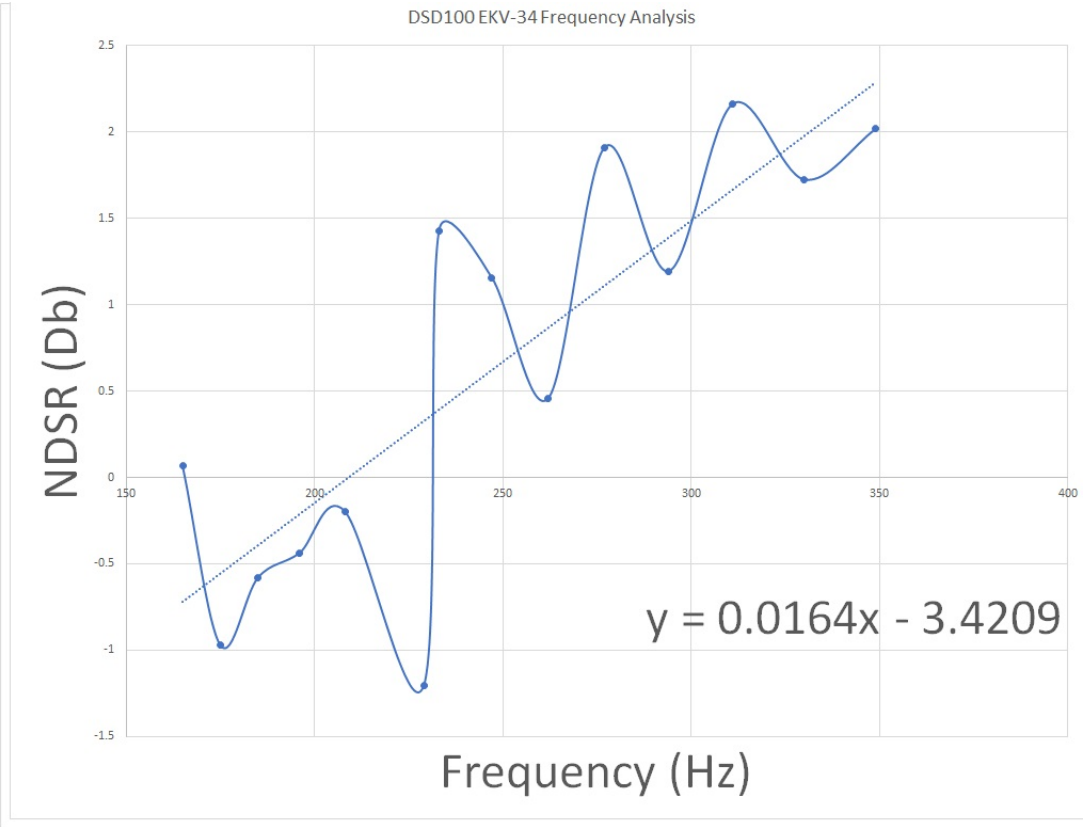

Figure 5.1: NSDR Graph for Frequency Analysis

difference in NSDR of 1.5Db on average. The slope of the accompaniment over 20 difference accompaniments from the DSD100 dataset is shown in table below.

5.3.1 MWV Testing

DSD100 Accompnaiment with MWV vocals	Slope of LSE NSDR
001 - ANiMAL - Clinic A	0.0175062662249768
002 - ANiMAL - Rockshow	0.0135507151628692
003 - Actions - One Minute Smile	0.014628225167387
004 - Al James - Schoolboy Facination	0.0172172063067036
005 - Angela Thomas Wade - Milk Cow Blues	0.0155333887485555
006 - Atlantis Bound - It Was My Fault For Waiting	0.0153118965390355
007 - BKS - Too Much	0.0182875321978301
008 - Bill Chudziak - Children Of No-one	0.0192272497216527
009 - Bobby Nobody - Stitch Up	0.0101307127747325
010 - Carlos Gonzalez - A Place For Us	0.0300621471097137
011 - Cnoc An Tursa - Bannockburn	0.0140314219002906
012 - Dark Ride - Burning Bridges	0.0121587469563734
013 - Drumtracks - Ghost Bitch	0.019024410058911
014 - Fergessen - Back From The Start	0.0145706175072749
015 - Fergessen - The Wind	0.0108563721915111
016 - Forkupines - Semantics	0.0144334817747698
017 - Girls Under Glass - We Feel Alright	0.0122316294565907
018 - Hollow Ground - Ill Fate	0.013514230862585
019 - James Elder _ Mark M Thompson - The English Actor	0.0171170161230281
020 - James May - Dont Let Go	0.0145751006657444

Table 5.1: Comparison of slopes for LSE line using MWV

The table above shows the DSD100 accompaniment that was combined with the MWV vocal dataset and the slope corresponding to the least squares line fit. While the slope of the lines are all quite small (<0.2) it is important to consider that the slope is in units of Db/Hz. Over the range of values in the dataset we can find the estimated difference between the Db NSDR level:

$$\text{Average slope} = 0.0157$$
$$\text{Range of Frequencies} = 165\text{Hz - } 350\text{Hz}$$

$$\texttt{Estimated difference in NSDR across Frequencies} = 0.0157 \times 350 - 165 = 2.9Db \tag{1}$$

This is a significant increase in effectiveness of RPCA across frequency. A 3Db increase in NSDR corresponds to approximately double the signal to distortion ratio.

5.3.2 EKV testing

Acompaniment from DSD100, mixed with EKV vocals	slope of LSE NSDR
001 - ANiMAL - Clinic A	0.0163575707365109
002 - ANiMAL - Rockshow	0.0232393580183783
003 - Actions - One Minute Smile	0.0203537210021792
004 - Al James - Schoolboy Facination	0.0220756964453295
005 - Angela Thomas Wade - Milk Cow Blues	0.0164097100733867
006 - Atlantis Bound - It Was My Fault For Waiting	0.0172263503216192
007 - BKS - Too Much	0.0313683456290835
008 - Bill Chudziak - Children Of No-one	0.0284547364962023
009 - Bobby Nobody - Stitch Up	0.0215925840801335
010 - Carlos Gonzalez - A Place For Us	0.0394097015310395
011 - Cnoc An Tursa - Bannockburn	0.0205175847868619
012 - Dark Ride - Burning Bridges	0.0236661922407127
013 - Drumtracks - Ghost Bitch	0.0189218761247329
014 - Fergessen - Back From The Start	0.0241208721814888
015 - Fergessen - The Wind	0.0173873044806292
016 - Forkupines - Semantics	0.0205764172543647
017 - Girls Under Glass - We Feel Alright	0.0189668856039586
018 - Hollow Ground - Ill Fate	0.0266832307331772
019 - James Elder _ Mark M Thompson - The English Actor	0.022075354197374
020 - James May - Dont Let Go	0.0209372778745099

Table 5.2: Comparison of slopes for LSE line using EKV

$$\text{Average slope} = 0.0225$$
$$\text{Range of Frequencies} = 165\text{Hz} - 350\text{Hz}$$

$$\texttt{Estimated difference in NSDR across Frequencies} = 0.0225 \times 350 - 165 = 4.1Db \tag{2}$$

Similar to the MWV vocals the EKV vocals also show a significant increase in effectiveness of RPCA for singing voice separation at higher frequencies of singing voice. An increase of

4.1Db over the range of frequencies tested corresponds to an increase of 2.5 in the NSDR as Db is a log scale.

5.4 Dicussion

The EKV and MWV dataset as the MIR-1K dataset and the DSD100 dataset use untrained and trained singers respectively. The results are consistant however as it is observable in both datasets there is a positive correlation between the increase in frequency and the NSDR level. This is strong evidence that the higher the singers vocal frequency range the better the outcome for RPCA for singing voice separation. The difference in Db levels across the different datasets is similar too which suggests that the increase in effectiveness is not due to the vocal training of the singers which recall is a difference between the MIR-1K and DSD100 dataset used in other sections of the report.

5.5 Impact

This could result in the RPCA for singing voice separation machine learning technique work better for female singers compared to male singers. It could also be attributed to the contributor having less variance in her voice at higher frequencies although this is not investigated.

In the context of RPCA across different genres this is an important result as only 22.4 percent of the billboard top 100 songs utilize female vocalists [1]. It is not necessarily that RPCA is a poor technique choice for singing voice separation when the vocalist is male but knowledge that it works better for higher frequency voices can help avoid biases in future work.

5.6 Further work

It is important to consider that the results may not be consistent across other datasets. Further work could include testing RPCA for frequencies with respect to male and female voices that span the same frequency range. One limitation of the experiment is the number of data points for frequency. It may be possible to manipulate the samples in more data by interpolating the frequencies of the accompaniment by either using microtonal singing or by artificially increasing singing frequencies and analysising their impact.

[1] https://www.nytimes.com/2018/01/25/arts/music/music-industry-gender-study-women-artists-producers.html

6 Overall Report Conclusions

In conclusion it appears that it is not trivial to select a parameter for K based on genre. The evidence suggests that for each dataset there are other factors that impact the effectiveness of RPCA for SVS such as the frequency of the singer (and hence possible the gender of the singer). It still may be the case that for a given genre RPCA can be improved but more refined and genre specific dataset would be needed in order to test such a hypothesis in more depth.

It should be noted that the lack of impact produced by trained and untrained singers was not the expected result and

6.1 Future Work

A more in depth analysis of genre is needed for singing voice separation techniques to determine if they are genre invariant. An obvious goal for the future would be developing more datasets that for singing voice separation as the current available datasets are inadequate for in depth testing of song characteristics.

Bibliography

[1] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 57–60.

[2] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[3] C. Févotte, R. Gribonval, and E. Vincent, "BSS$_E$*VALToolboxUserGuide* $-$ $-$*Revision*2.0," 012005.

[4] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent f0 estimation and source separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 574–578.

[5] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 106–110.