Master in Sound and Music Computing

Universitat Pompeu Fabra

Vocal Source Separation for Carnatic Music

Adithi Shankar Sivasankar

Supervisors: Prof. Xavier Serra i Casals and Genís Plaja i Roglans

August 2023

Master in Sound and Music Computing

Universitat Pompeu Fabra

Vocal Source Separation for Carnatic Music

Adithi Shankar Sivasankar

Supervisors: Prof. Xavier Serra i Casals and Genís Plaja i Roglans

August 2023

Contents

1 Introduction **1**

1.1 Context 1

1.2 Motivation 4

1.3 Objectives 5

2 State of the Art **7**

2.1 CompMusic Project 7

2.1.1 Curated corpora and Research as part of the CompMusic Project . . . 8

2.1.2 Riyaz App and the relevance of Source Separation 9

2.2 Multi track Datasets used for Source Separation 10

2.2.1 MedleyDB 10

2.2.2 MUSDB18 11

2.3 Types of Source Separation Models 11

2.3.1 Waveform and Spectrogram based models 11

2.3.2 Spleeter by Deezer 12

2.3.3 Hybrid Demucs 12

2.4 The Music Demixing Challenge 14

3 A Carnatic Multi-Track Dataset **15**

3.1 Biases in Existing Multitrack datasets 15

3.2 Stems in Carnatic Music 16

3.2.1 The Importance of the Singing Voice 17

3.2.2 Accompanying melodic instruments 18

3.2.3 Percussion instruments . . . 18
3.2.4 The Tanpura . . . 19
3.3 The Recording process . . . 20
3.3.1 Singers and the List of Compositions . . . 20
3.3.2 Initial setbacks in the process: Issues with Reverb . . . 22
3.4 Data Augmentation and Credibility related concerns . . . 22
3.4.1 The Mixing Process . . . 24
3.4.2 Challenges and Creative choices in the process . . . 24

4 The Retraining and Fine-tuning Process 26
4.1 The Model and Prerequisites for the Retraining process . . . 26
4.2 Model Retraining and fine-tuning process . . . 28
4.3 Retraining and Fine-tuning Experiments . . . 29
4.3.1 Loading and Validating the model . . . 30
4.4 Conclusions . . . 31

5 Evaluation and Results 32
5.1 The Source to Distortion Ratio . . . 33
5.2 Results from Model 1 . . . 33
5.3 Results from Model 2 . . . 35

6 Conclusions and Future Work 36
6.1 Summary of the thesis . . . 36
6.2 Extracted Conclusions . . . 37
6.3 Possible research directions . . . 38

List of Figures 40

List of Tables 41

Bibliography 42

Acknowledgement

This year has been a tremendous change in my life and I have multiple people to thank for helping get through. I'm extremely grateful to my parents for letting me pursue what my heart desires and move to Barcelona. All of this would not have been possible if not for Prof. Xavier Serra. I'm thankful to him for introducing me to the world of Music Information Retrieval and giving me direction to take forward my dreams. Genis Plaja Roglans, my supervisor and guide has been the most wonderful influence this year, inspiring and helping me in every step of the thesis and research in general. Special thanks to the team of Riyaz; Gopal for being the firm yet gentle mentor and Aditi for being the reassuring voice that I most needed when I did not believe in myself.

I have to thank my friends at the MTG; Dilip, Jyoti, Tom and Ahana for motivating and helping me push to the best of my abilities. I'm also thankful to my SMC peers for being extremely supportive throughout this intense academic year. I'm very grateful to Ameena aunty for being home away from home, and also to Ayushi, Dayal and Avinash uncle for always being there for me . I also have to thank my friends back home; Gopesh, Lakshana, Chaitanya and Algates for always having my back.

Finally, I have to thank my mother for introducing me to the world of Carnatic Music and if she had not persevered, I would not have been any bit capable of doing anything that I'm doing today.

Thank you everyone,

Adithi

Abstract

Carnatic Music is a Classical music form that originates from the South of India and is extremely varied from Western genres. Music Information Retrieval (MIR) has predominantly been used to tackle problems in western musical genres and cannot be adapted to non western musical styles like Carnatic Music due to the fundamental difference in melody, rhythm, instrumentation, nature of compositions and improvisations. Due to these conceptual differences emerged MIR tasks specific for the use case of Carnatic Music. Researchers have constantly been using domain knowledge and technology driven ideas to tackle tasks like Melodic analysis, Rhythmic analysis and Structural segmentation. Melodic analysis of Carnatic Music has been a cornerstone in MIR research and heavily relies on the singing voice because the singer offers the main melody. The problem is that the singing voice is not isolated and has melodic, percussion and drone instruments as accompaniment. Separating the singing voice from the accompanying instruments usually comes with issues like bleeding of the accompanying instruments and loss of melodic information. This in turn has an adverse effect on the melodic analysis. The datasets used for Carnatic-MIR are concert recordings of different artistes with accompanying instruments and there is a lack of clean isolated singing voice tracks. Existing Source Separation models are trained extensively on multi-track audio of the rock and pop genre and do not generalize well for the use case of Carnatic music. How do we improve Singing Voice Source Separation for Carnatic Music given the above constraints? In this work, the possible contributions to mitigate the existing issue are ; 1) Creating a dataset of isolated Carnatic music stems. 2) Reusing multi-track audio with bleeding from the Saraga dataset. 3) Retraining and fine tuning existing State of the art Source Separation models. We hope that this effort to improve Source Separation for Carnatic Music can help overcome existing shortcomings and generalize well for Carnatic music datasets in the literature and in turn improve melodic analysis of this music culture.

Keywords: Carnatic Music; Source Separation, Melodic analysis.

Chapter 1

Introduction

This section introduces fundamental ideas that the thesis is built on top of. The discussions include musicological insights into Carnatic Music, Music Information Retrieval (MIR) technologies and Computational tools that help analysis of Carnatic Music, shortcomings in the analysis of the art form, the problem of Source Separation and how it is different for Carnatic music, how does it affect melodic analysis and how can it be approached. The section also gives a brief overview of the melodic analysis trends in Carnatic MIR and the role of low-level feature extraction.

1.1 Context

Music Information Retrieval(MIR) is the interdisciplinary science of retrieving information from Music. It brings together the understanding of musicology, perception, cognition and technology to solve tasks like Audio Classification, Source Separation, Automatic Music Transcription, Optical Music Recognition and many others. Previously, the field of MIR heavily relied on Signal processing methods for most of the above-mentioned tasks but the advent of Machine Learning (ML) and Deep Learning (DL) technologies has considerably reduced the use of Signal Processing. As summarized by Casey et al. [1], content-based MIR and data-driven methods started gaining more attention. However, the issue of data scarcity arises, given that most Deep Learning models need large amounts of data to be trained with. To ad-

dress this issue, task-specific datasets were painstakingly curated and open-sourced for research purposes. Given the popularity, most of the tasks were addressed to the use case of Western music and the datasets were of the same genre too.

Carnatic Music, a classical music art form from the South of India garnered recognition from the MIR community through the CompMusic project initiated by Prof Xavier Serra at the Music Technology Group. The project was conceptualized after realising the need for a multicultural approach in MIR research [2]. Carnatic Music was one among the five other non-western musical traditions that was researched as part of the project. These art forms have tradition-specific musical characteristics that cannot be retrieved using technologies built for the Western Genre. Researchers worked extensively on building a corpus to facilitate MIR research in these five traditions. Out of the five traditions, Indian Art Music comprising Carnatic and Hindustani Music, was researched extensively and continues to be researched.

The research has been broadly divided into three parts; Melodic analysis, Rhythmic analysis and Structural analysis. Melody in Carnatic Music is heavily embellished with ornamentation called gamakas and the notes do not correspond to a stable pitch position due to the nature of the art form. These pitch positions are collectively defined as svarasthanas [3]. Carnatic Music also has the concept of Raga which essentially is the scale, with gamakas and characteristic motifs defining it. These tradition-informed concepts gave researchers the scope to work on tasks like Automatic Raga recognition, Melodic motif mining, Intonation analysis and some others. The instrumentation in Carnatic Music has the vocalist who usually offers the main melody, an accompanying violinist who layers the melody on the top of the vocals, the mridangam that offers the main percussion support and the ghatam and kanjira that offers the supporting percussion. The tanpura is the drone instrument that offers the tonic of the singer. All these instruments come together coherently as part of the concert.

Given that most Carnatic music concerts are vocalist-driven, we depend on the vocal source for any melodic analysis task. The datasets curated for Carnatic music research are sourced from concerts that are recorded live. There is a lack of clean

vocal stem data and we depend on Source separation models to obtain vocal stems from datasets.

State of the art Source Separation models are majorly Deep Learning based models trained on datasets like MUSDB18 [4] and MedleyDB [5] that are mainly of the rock and pop genre. These models are trained with instrument stems like the vocals, bass, drums and others. Most commonly used Source Separation models include Spleeter [6], Demucs [7] and Open-Unmix [8]. The process of Source separation is demonstrated in Figure 1.

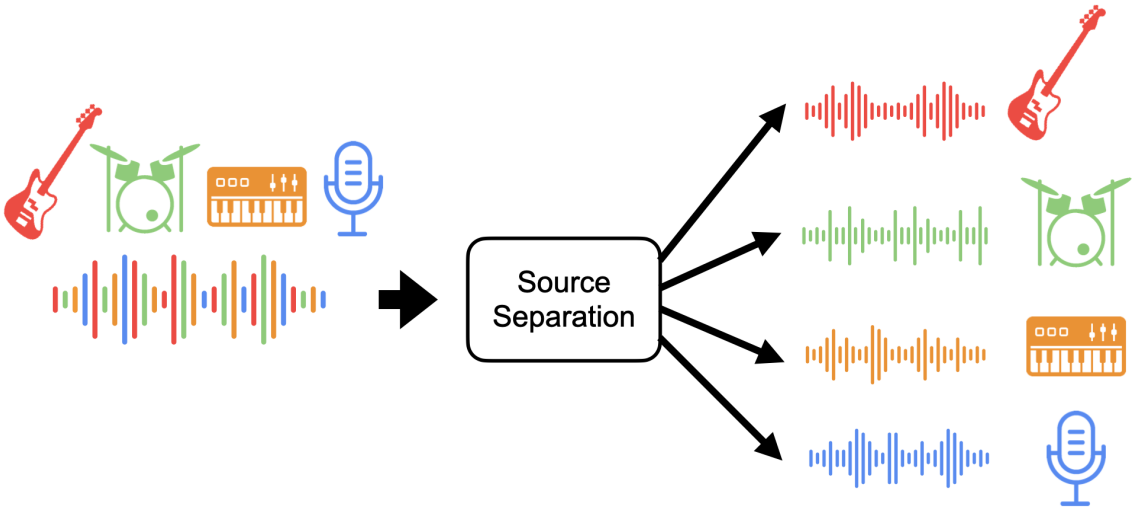

Figure 1: The Process of Music Source Separation

When used to separate Carnatic Music data, they do not generalize well due to the presence of unfamiliar stems like the pitched percussion, drone and violin. There is a severe bleed of these instruments in the separated vocals which can affect the overall melodic analysis pipeline. The types of bleed that are evident when using spleeter are:

1) The violin is not removed, and given that it overlaps the vocals with a slight lag the model assumes it to be part of the vocals.

2) Mridangam is a pitched percussion instrument and spleeter does not identify pitched percussion well enough.

3) The drone is not completely removed.

From the above analysis, it is understood that the models are not familiar with Carnatic music stems and does not generalise well for the use case. With the given background, the research question that we are aiming to address is, **Could we reuse existing data or build a new multitrack Carnatic music dataset to retrain or finetune existing state-of-the-art Source separation models to improve Source separation for the use case of Carnatic music? Is it feasible to use data with bleeding to retrain or finetune models and improve Source separation?**

1.2 Motivation

How can the bleed interfere with the melody? The instruments that are part of the bleeding are all pitched instruments and these can interfere with the actual fundamental frequency causing undesirable frequencies in the output. Given that the melodic analysis rests on the Separated vocals, improving Source Separation for Carnatic music is a task that has been emphasized by several researchers within the community [9].

From the above discussions, it is evident that improving Source Separation for Carnatic music will have positive implications in the melodic pipeline. In the literature, Source Separation models have always been trained with multi-track data and there is a lack of multi-track data for Carnatic music. Saraga Carnatic [10] is the most widely used dataset for MIR research in Carnatic Music. It is a 60-hour-long dataset with several concert recordings of different artists. The dataset also includes annotations and metadata relevant to each piece that is part of the concert recordings. The annotations include the fundamental frequency, tonic and sama detection. the metadata has relevant information like the name of the raga, the section timestamps denoting the alapana, composition, kalpana swara or neraval, the duration of the composition and others.

Saraga Carnatic has about 150 recordings that have multi-track audio with leakage. These multi-track audio recordings are recordings of separate stems but with leakage

because the stems were not isolated as they are recorded as part of the concert. While analysing the Spleeter Source Separated Carnatic vocal stem against the multi-track vocal stem in Saraga Carnatic, they show similar patterns of leakage. Using just the multi-tracks with leakage will not help the model understand the above-listed issues with Spleeter.

Wrapping up the discussion, we conclude that a data-driven DL approach can be efficiently used to solve the problem of Source Separation. What lies ahead is how to create new datasets that can help with the task and also reuse data to facilitate it.

1.3 Objectives

In this section, we define the objectives and flow for the upcoming work. This approach is based on a two-step methodology that involves a data creation process and a Deep Learning training process. We also discuss the metrics to evaluate the inferences obtained.

The task can be approached with a two-step methodology: 1.1) Creating a Carnatic music multi-track dataset similar to MedleyDB and MUSDB18. 1.2) Creating a multi-track dataset with Saraga carnatic data with bleeding and 2) Retraining or Fine-tuning existing State of the Art with the above datasets.

Given that our main concentration is on improving Singing Voice Source Separation, we chose a 2 stem (Vocals and Accompaniment) model to retrain or finetune. The most commonly used model for Carnatic music Source Separation is Spleeter and this model is chosen for retraining or finetuning. The dataset is also mixed keeping in mind the 2 stem model. The Accompaniment stem of the dataset is the sum of the mridangam, violin and tanpura stems and the Vocal stem remains as such. The dataset also has the mix stem which is the sum of time-aligned Vocal and Accompaniment stems.

Another objective of the thesis is to reuse data from existing datasets like Saraga Carnatic. These datasets are retrieved from standard concert recordings and mirror

the concert format precisely. Saraga Carnatic has large amounts of data with bleed that can be valuable if used efficiently. Saraga Carnatic also has a wide variety of accompaniment instruments like the ghatam, morsing and kanjira which can be used to improve Source Separation.

The most commonly used Evaluation metric for Source Separation is the Source to Distortion Ratio (SDR). SDR is usually considered to be an overall measure of how good a source sounds. Though SDR is used widely as the evaluation metric, it is argued that it is not the most holistic metric for the evaluation of Source-separated stems [11]. The evaluation will be a comparison between the existing 2 stem Spleeter and the retrained/finetuned Carnatic custom Spleeter models.

While evaluating Saraga Carnatic Vocal stems, there is a lack of clean ground truth. So a general perceptual comparison between the 2-stem Spleeter and custom Carnatic Spleeter models in terms of bleeding of the different instruments can used as a starting point for evaluation.

Chapter 2

State of the Art

This section introduces the State of the Art in Carnatic-MIR and Source Separation. There is a briefing about the CompMusic project that was instrumental in taking forward MIR research in non-Western classical music. There is a dedicated section to elucidate the architecture of Source Separation models and their application to Carnatic-MIR. This includes a small introduction to the Music Demixing challenge. There are also pointers to possible future research in the field of Carnatic-MIR.

2.1 CompMusic Project

CompMusic (Computational Models for the discovery of the world's music) is a large project funded by the European Research Council and coordinated by Prof. Xavier Serra from the Music Technology Group of the Universitat Pompeu Fabra in Barcelona. It aimed at building tradition-informed Computational tools for the five chosen non-western classical music forms namely Carnatic Music, Hindustani Music, Turkish Makam, Arab Andalusian music, and Beijing Opera. These forms are extremely different from what Western musical forms are and require to be addressed exclusively through musicological knowledge and technological methods. The project spanned between 2011 to 2017 and several researchers worked on addressing these five musical forms from a multicultural perspective to build bespoke tools through analysis of melody, rhythm and structure.

Building Computational tools required data from these traditions and researchers meticulously curated datasets that can be generalisable for multiple tasks in the same genre [12]. These datasets were built from different sources like concert recordings, and personal live recordings by artists and were also open-sourced to facilitate further research. Section 2.1.1 explains the datasets and the research in Indian Art Music. Section 2.1.2 talks about the Riyaz app and the need for improving Source Separation from the perspective of a company.

2.1.1 Curated corpora and Research as part of the CompMusic Project

The research corpora built for Carnatic Music played a major role in identifying different dimensions in Carnatic-MIR. The Carnatic corpus [13] includes 2,380 audio recordings (235 concerts, 500 hours), covering 259 artists, 965 compositions, 227 ragas, 15 talas, and 15 forms. Below is the list of datasets that were curated for Carnatic Music research. These datasets are mainly divided into two tasks in Carnatic MIR; Melodic analysis and Rhythmic analysis. The Indian Music Tonic Dataset comprises 597 commercially available audio music recordings of Indian art music (Hindustani and Carnatic music), each manually annotated with the tonic of the lead artist. This dataset is used as the test corpus for the development of tonic identification approaches. The Carnatic varnam dataset is a collection of 28 solo vocal recordings, recorded for our research on intonation analysis of Carnatic ragas. The collection consists of audio recordings, time-aligned tala cycle annotations and swara notations in a machine-readable format.

The Carnatic Music Rhythm Dataset is a sub-collection of 176 excerpts (16.6 hours) in four taalas of Carnatic music with audio, associated tala-related metadata and time-aligned markers indicating the progression through the tala cycles. It is useful as a test corpus for many automatic rhythm analysis tasks in Carnatic music. A subset with 118 two-minute-long excerpts (about 4 hours) is also available with equivalent content. The Mridangam Stroke dataset is a collection of 7162 audio examples of individual strokes of the Mridangam in various tonics. The dataset com-

prises 10 different strokes played on Mridangams with 6 different tonic values. The dataset can be used for training models for each Mridangam stroke. The Mridangam Tani-avarthanam dataset is a transcribed collection of two tani-avarthanams played by the renowned Mridangam maestro Padmavibhushan Umayalpuram K. Sivaraman. The audio was recorded at IIT Madras, India and annotated by professional Carnatic percussionists. It consists of about 24 min of audio and 8800 strokes.

These datasets have been widely used in tasks in Carnatic-MIR. Summarising some of the research using these datasets, the Carnatic Varnam dataset was used by GK Koduri et al. [14] for the task of intonation analysis, the Carnatic Rhythm dataset was used for beat tracking and onset detection tasks in Indian Art music [15, 16] and the Indian Music Tonic dataset was used for tonic identification using multiple approaches [17, 18, 19].

2.1.2 Riyaz App and the relevance of Source Separation

The Riyaz app, an Indian Classical music education app was conceptualized as a result of years of consistent research as part of the CompMusic. The app was co-founded by Dr Sankalp Gulati and Dr Gopala Krishna Koduri who were both researchers, active in the field of MIR for Indian Art Music. The app is designed in a way that the user can sing a song of their choice and the singing voice is compared against the fundamental frequency of the original song. The app gives pointers when the fundamental frequencies don't match with some given leniency. It has hundreds of recorded lessons from beginner to advanced levels. It was also recently expanded to accommodate Indian film music and semi-classical devotional music.

The pipeline for Fundamental frequency extraction is a multi-step process. The initial step would be Source Separating the vocals from the songs as most available songs are the mastered mix. Due to the fact that the app is designed for singers, the source-separated vocals become the source of data against which the singer will be evaluated. As discussed in Section 1, source separation for Indian Art music has induced bleeding due to the instrumentation. Data with bleeding can alter the fundamental frequency and induce undesirable frequencies and onsets. The problem

of Source Separation becomes relevant to the use case of the app here.

This thesis is a collaboration with the Riyaz app to mitigate the issue of Source Separation, and can also be beneficial to improve the melodic analysis of Carnatic Music. Carnatic music is given emphasis here since the app houses multiple Carnatic music lessons and compositions and aims to improve the machine-assisted music education space for Carnatic music. Major issues that affect the app due to bleed in the Source Separated vocals are the unexpected attacks from the percussion instruments that clash with the vocals, and the violin and tanpura bleed which alter the melody. The issue remains consistent for both the app and the case of Carnatic MIR.

2.2 Multi track Datasets used for Source Separation

This section lists out the most commonly used multi-track datasets for training Source Separation algorithms. The size, encoding format, source and type of stems are elaborated below. The datasets are ordered by the date of release.

2.2.1 MedleyDB

MedleyDB is a dataset with 122 multi-tracks with melody annotations [5]. The tracks are sourced from different sources. Independent Artists contributed 30 songs, NYU's Dolan Recording Studio contributed 32 songs, Weathervane Music contributed 25 songs, and Music Delta contributed 35 songs, summing up to a total of 122 songs. All tracks are encoded at 44.1 KHz the mixed tracks are stereo and the stems are mono. The audio files are in .wav format. Each track has metadata available in a YAML file denoting details like the artist, composer, instrumentation and others. The tracks also come with fundamental frequency annotations and instrument activations. MedleyDB was extended to MedleyDB2.0 with an addition of 74 tracks making a total of 196 tracks.

2.2.2 MUSDB18

The MUSDB18 dataset has a total of 150 tracks [4]. The dataset has 4 stems namely Vocals, Drums, Bass and Others. All signals are stereo and encoded at 44.1KHz. The dataset is divided into Train and Test with 100 and 50 tracks respectively. The data from MUSDB is composed of data from different sources.100 tracks are taken from the DSD100 dataset, which is itself derived from The 'Mixing Secrets' Free Multitrack Download Library. 46 tracks are taken from the MedleyDB dataset. 2 tracks were kindly provided by Native Instruments originally part of their stems pack.2 tracks are from the Canadian rock band The Easton Ellises as part of the Heise stems remix competition. The dataset is almost 10 hours long and has two formats. The uncompressed MUSDB18-HQ and the compressed stems MUSDB18. The compressed stem format has the stems encoded in the Native instruments stem format.

2.3 Types of Source Separation Models

This section briefs the most commonly used Source Separation algorithms in the literature. The architecture of each of these models is described briefly and a short note on what kind of bleed is induced by these models when used to Source Separate Carnatic Music is written. There is also a brief note on Waveform-based models and Spectrogram-based models that form the basis of these models in the literature.

2.3.1 Waveform and Spectrogram based models

Waveform-based models operate directly on the raw audio waveform data. These models typically use deep learning architectures to learn the complex relationships between mixed waveforms and individual sources. Several models use waveform data for Source separation [20, 21, 22]. The idea is to directly estimate the waveform of each source from the mixed waveform. Spectrogram-based models work in the time-frequency domain by converting the audio waveform into a spectrogram representation. A spectrogram is a visual representation that displays the frequency

content of a signal over time. Spectrogram-based models aim to use the spectrogram to isolate individual sources.

Both these models require a sizeable amount of data to train with, but Waveform-based models require considerably more data because raw audio is generally more intricate having nuanced temporal information. This is the very reason that makes models Computationally intensive. On the other hand, Spectrogram based models capture the frequency over time making it a concise representation of the signal. Hence, it requires considerably low computation and data. Current state-of-the-art models are mainly spectrogram-based [23, 24], or hybrid models that leverage both domains [25].

2.3.2 Spleeter by Deezer

Spleeter is a spectrogram-based Source Separation model developed by Deezer Paris. Spleeter contains pre-trained models for 2 stem separation - Vocals and Accompaniment, 4 stems separation - Vocals, bass, drums and other, and 5 stem separation - Vocals, bass, drums, piano and other. Spleeter is one of the most efficient and fastest models available for source separation. It is the most commonly used Source Separation model for Carnatic-MIR. Spleeter is based on the U-Net architecture. The U-net is an encoder/decoder Convolutional Neural Network (CNN) architecture with skip connections [26]. Figure 2 demonstrates the architecture of a U-Net network. Spleeter uses a 12-layer U-net (6 layers for the encoder and 6 for the decoder). Spleeter can separate the entire MUSDB18 dataset(3 hours and 27 minutes long) into 4 stems in less than 2 minutes. The Signal to Distortion Ratio(SDR) of Vocals in Spleeter Multi-Channel Wiener Filtering mode is 6.85 dB which is on par with Open-Unmix and Demucs.

2.3.3 Hybrid Demucs

Hybrid Demucs is a Hybrid-Spectogram and Waveform-based model that won the Music Demixing Challenge 2021. Hybrid Demucs use the Original Demucs architecture and extend it. The model is composed of a temporal branch, a spectral branch,

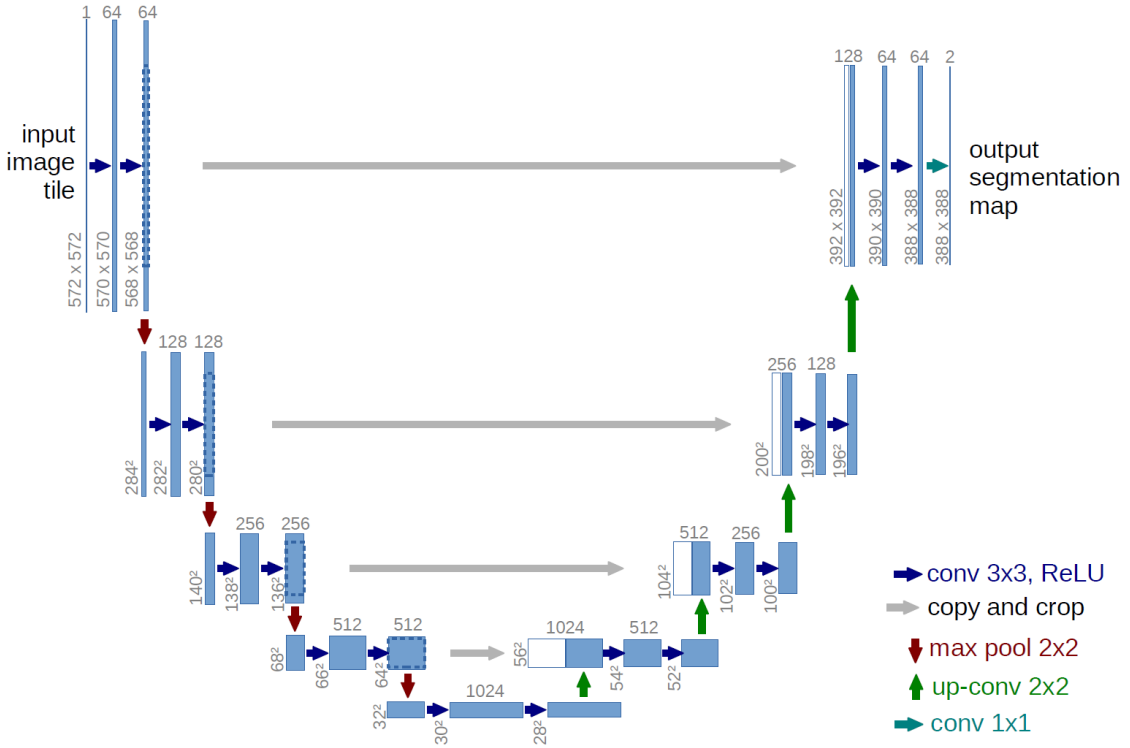

Figure 2: The architecture of a U-Net

and shared layers. The temporal branch takes the input waveform and processes it like the standard Demucs. It contains 5 layers, which are going to reduce the number of time steps by a factor of 5. Compared with the original architecture, all ReLU activations are replaced by Gaussian Error Linear Units (GELU). The model is pre-trained with several datasets including MusDB. The winning model mdx was trained on MUSDB HQ. The mdx model separates the tracks into 4 stems- Vocals, Drums, Bass and Other. The Vocal SDR of Demucs is 7.97, and it performs better than Spleeter in this case.

In spite of having a better metric than Spleeter, Spleeter continues to be used for Source Separation in most Carnatic-MIR research [27, 28, 29]. It is because of a characteristic of the tradition that Demucs fails to cope with. Carnatic music has the vocals and violin layering the vocals with a slight lag. There are instances where the violin is in complete unison with the vocals and the violin sounds more dominant. Demucs identifies these portions as accompaniment and removes the vocals completely. This leads to a complete loss of data and causes an abrupt

silence in the melody.

2.4 The Music Demixing Challenge

The Music Demixing Challenge is a contest in the field of MIR that focuses on the task of Source Separation from music signals [30]. It specifically aims at separating musical signals into individual stems- Vocals, bass, drums and others. The MUSDB18 dataset is chosen to evaluate and benchmark Source separation models. The metrics used for evaluation are Signal to Distortion Ratio(SDR), Signal to Inference ratio (SIR) and Signal to Artifact ratio (SAR).

For the competition, contestants have to submit their Source separation models which will be evaluated over a hidden test set. The winning models will be chosen on the basis of the above-mentioned evaluation metrics. The Hybrid Demucs mdx model was one of the models that was on the leaderboard of the Demixing challenge.

Chapter 3

A Carnatic Multi-Track Dataset

Creating a Multitrack dataset for Carnatic music has multiple layers. In this section, the creative process is explained and the choices are justified. The creative process mainly includes the choice of stems, compositions, raga, and the techniques used for augmentation. The recording process is explained in detail and the mixing process is also highlighted. Issues related to biases in existing datasets and the credibility of augmented data are discussed. The challenges faced during the remote recording process are punctuated.

3.1 Biases in Existing Multitrack datasets

The datasets discussed in Chapter 2 are large datasets inclined towards the Western Pop Genre. The stems include the Vocals, Drums, Bass and Others. The Drums and Bass are predominantly used in Western genres and do not feature in most music from other parts of the world. This makes generalizing for other musical forms, not a feasible enough task. As summarized by Holzapfel et al [31], there is a bias towards "US popular music or a restricted part of European ‘Classical concert music’".To the best of our knowledge, there are no multitrack datasets existing for non-western classical music forms, but there are several instances of non-western musical research using Source Separation models trained with these datasets. This emphasizes the need to mitigate the bias and have models that can generalize better

for tradition-specific instruments.

Carnatic Music being one of the highly researched non-western classical music forms in MIR struggles with these biases due to the specificity of the genre. Building a Carnatic music-specific multitrack dataset could possibly help alleviate these biases and generalize better for the art form.

3.2 Stems in Carnatic Music

Carnatic Music is mainly performed in the concert format. Figure 3 depicts a Carnatic Music concert setting. There are two lead vocalists (Generally there is one), a violinist, a mridangist and a person playing the tanpura sitting right behind the vocalist. The vocalist assumes an exalted position in most concerts with them deciding the choice of compositions, length of the alapana (melodic improvisation around a raga), the refrain around which the kalpana svara (melodic rhythmic improvisation) is sung and others. The violinist supports the vocalist to give a harmonious output. The mridangist is the main source of percussion and keeps rhythm and layers the melody. The ghatam, kanjira and morsing are the supporting percussion instruments that are usually featured in concerts.

Out of the percussion instruments, the mridangam has the capability to produce multiple timbres due to its construction. The mridangam is tuned to the tonic making it a pitched percussion instrument. Multiple different strokes played in the mridangam can translate to multiple variations in sound. This makes the mridangam one of the most unique components of a Carnatic music concert.

Keeping the instrumentation in mind, a dataset with two stems- Vocals and Accompaniment is a seemingly ideal choice. The Accompaniment stem will include the time-aligned mix of the Mridangam, Violin and Tanpura. The Ghatam, Kanjira and morsing are instruments with lesser timbral variation as compared to the mridangam and are supporting instruments. So the mridangam is taken as the percussion-providing source.

Figure 3: A Carnatic Music Concert

3.2.1 The Importance of the Singing Voice

In a vocalist-led Carnatic Music concert, the concert is singing-driven. Meaning that most of the creative choices regarding the melody of the concert are made by the vocalist. A concert comprises a series of multiple compositions and improvisatory pieces in different ragas and taalas and the content of the concert is decided by the singer. From an MIR perspective this implies that for any melodic analysis task, we rely on the singing voice.

Even for a task like Structural segmentation, the vocals remain a main source of information due to the fact that the structure of a concert is partially orchestrated by the vocalist(alongside changes in accompaniment instruments). The dependence on the vocals for relevant information of the tradition for analysis makes Singing

voice Source separation a pertinent task.

3.2.2 Accompanying melodic instruments

In earlier times, the vocalist was accompanied by instruments like the Veena and the Nagaswaram. The Nagaswaram has a robust sound and given that microphones were not used back then, it could almost drown the singing voice. Subsequently, the veena started being used to accompany the vocals. The veena has a sweet and mellow sound but is not versatile for all pitches due to its construction.

The violin entered the Carnatic music space in the late 17th century. It is an import from the West and was adapted to playing Carnatic music. The size made it convenient to travel with and it could be adapted to playing different tonic pitches, unlike the veena. To suit the Carnatic style, several tuning combinations and playing techniques were experimented with. The violin that we see in Carnatic Music concerts today is tailored to suit the tradition.

3.2.3 Percussion instruments

The Mridangam is a two-headed drum with a hollow body made of jackfruit tree wood and membranes made with cow skin. The left and right membranes are used by the artiste to generate varied strokes to match the rhythmic solfege in Carnatic music called Konakkol. The ghatam is a clay pot with a narrow mouth designed with a specific thickness to produce the desired tonal quality. Striking different parts of the ghatam with different parts of the hand can produce different tones. The ghatam is often times paired with the mridangam in concerts as an accompanying percussion instrument.

The Kanjira is a South Indian frame drum, with a circular wooden frame wrapped with monitor lizard skin on one side. The instrument is commonly used in North Indian folk music and was adapted to Carnatic music in the late 19th century. The instrument is not pitched and remains the same for any tonic. The morsing is the south Indian jaw harp that follows the konakkol pattern to shadow the mridangam

in concerts. It is also widely used in North Indian folk music.

3.2.4 The Tanpura

The Tanpura or Tambura in Carnatic Music is the drone instrument offering the tonic of the vocalist in every concert. It provides a continuous harmonic drone and helps sustain the melody of another instrument or singer. It is played by plucking a cycle of four strings in a continuous fashion throughout the concert providing a canvas for the concert. The tanpura is pitch specific and comes in various pitches. The male pitches have a larger resonating chamber and the female pitches have a smaller resonating chamber. The Tanpura is considerably big and fragile and transporting it requires great care. A digital tanpura box (also called Shruti box), which played the synthesized sound of a tanpura in the desired pitch, was introduced to help with the size constraint. The tanpura was also made into different apps that could be used on mobile phones and laptops. The digital box is an alternative that has helped several musicians and students but has never been able to match the resonance of a real tanpura.

3.3 The Recording process

The Recording process is briefed in this section. The recording section starts with choosing the singers and a list of compositions that are commonly defined throughout Carnatic Music. The setbacks during a remote recording process are identified and discussed. The creative liberties taken as part of the process are listed and justified.

The intention of creating a Carnatic music-specific dataset is to improve Source Separation for Carnatic Music. We should note here that Carnatic music datasets that are part of the literature are live recordings in concert settings. The multitrack data that we have as part of the Saraga dataset could possibly used for finetuning, but the data contains severe leakage from the other sources It would not be ideal to just use this data for the training process. The dataset that we are creating aims to have clean Vocal and Accompaniment stems free from bleeding for the model to learn better.

The data that we are creating are recorded separately by different artists from different places in the world with bare minimum equipment. The dataset might not mirror the quality of the State-of-the-art datasets but attempts to inform the model of the kind of instrumentation used in the tradition to improve Source Separation.

3.3.1 Singers and the List of Compositions

Three singers, two male and one female are recorded for the dataset. The singers are recorded in their respective pitches: C Sharp, D and G. All three singers have been formally trained in Carnatic music for at least 12-15 years under different teachers. One thing to note is that in Carnatic music, there are different schools of thought or Bani as it is called and within these schools, there are slight changes in the way compositions are rendered [32]. These chosen singers have trained under different teachers and the way they render compositions might differ considerably. So Varnams were chosen for the singers to record. The reason for choosing Varnams is because they are learnt through standardized notations from books and the repetitions stay uniform.

Other than the Varnams, a list of four kritis are chosen. Every kriti has a structure starting with the Pallavi, and then the Anupallavi and finally the Charanam. Each of the segments is sung in multiple melodic ways within the raga grammar, called the Sangati and is repeated once or twice. So a list of the sangatis and its respective repetitions are made for the singers to adhere to. The singers are uniformly asked to record at a tempo of 60 Beats Per Minute. The idea of keeping the singing uniform is to record just one set of recordings of the Violin and Mridangam and use augmenting techniques to pitch shift these stems to juxtapose the recordings of the other two singers. The augmentation techniques are discussed in detail in section 3.4.

The reason for recording kritis in spite of the non-uniformity is to have impromptu variation in melody that the violin adapts and layers with a lag [33]. Existing Source Separation models fail to cope with this nature of the tradition and induce bleeding into the separated vocals.

The dataset has compositions in eight different ragas and two taalas. The list of compositions with the respective raga and taala are listed in Table 1.

Table 1: List of Compositions with its respective raga and taala

Name of Composition	**Raga**	**Taala**
Ninnukori (Varnam)	Mohanam	Adi (2 kalai)
Vanajakshiro (Varnam)	Kalyani	Adi (2 kalai)
Sami Ninne (Varnam)	Shri	Adi (2 kalai)
Valachi Vachi (Varnam)	Navaragamalika	Adi (2 kalai)
Mahaganapathim	Naattai	Adi
Maithreem Bhajatha	Yamuna Kalyani	Adi
Samajavaragamana	Hindolam	Adi
Bantureeti Kolu	Hamsanadam	Adi

3.3.2 Initial setbacks in the process: Issues with Reverb

The singers are asked to record on their phones hearing the Tanpura and metronome over their earphones. Initially, the male singer with pitch C Sharp is recorded. The singer recorded in his room with the Dolby Atmos app on his phone. Simultaneously, the Mridangist and Violinst are asked to record in their individual rooms, with the singer's recordings as reference. Due to the fact that the recordings are recorded without any equipment in small closed rooms, there is a layer of reverb over the recordings. When the separate stems are mixed, the layer of reverb gets multiplied and sounds extremely roomy. To soften the reverb, the above data is run through a de-reverb plug-in. The de-reverbed recordings have a slight change in timbre but have far less reverb than the actual recording. The reverbed data is treated as augmented data and kept as part of the dataset to have instances of different types of data [34].

The second singer recorded in the pitch of D The singer is asked to record in a larger room with the phone slightly far away from her. This reduced the reverb considerably. The third singer is asked to record with a microphone and the reverb is mitigated in most compositions. There are slight instances of reverb in some cases and it can be assumed that during these instances, the singer stayed close to the microphone.

Table 2 describes the list of compositions, information if the singers have recorded the respective song and the duration of the song in seconds.

3.4 Data Augmentation and Credibility related concerns

Data Augmentation can be divided into two types; Intra and Inter augmentation. Intra augmentation here refers to the pitch shifting of the Violin and the Mridangam to juxtapose the recordings of the other two singing voices recorded in different pitches. Inter Augmentation refers to the complete augmentation of all the stems of

Table 2: List of Compositions with singer-related information and duration

Composition	**Singer 1(C Sharp)**	**Singer 2 (D)**	**Singer 3 (G)**	**Duration**
Ninnukori (Varnam)	Yes	Yes	Yes	525.854
Vanajakshiro (Varnam)	Yes	Yes	Yes	568.085
Sami Ninne (Varnam)	Yes	Yes	Yes	676.082
Valachi Vachi (Varnam)	Yes	Yes	No	703.362
Mahaganapathim	Yes	Yes	Yes	264.84
Maithreem Bhajatha	Yes	Yes	Yes	191.957
Samajavaragamana	Yes	No	No	266.019
Bantureeti Kolu	Yes	Yes	Yes	306.004

a respective track. Both these techniques of augmentation are part of the creative choices and were made taking into account several factors.

For Intra Augmentation, we know that the violin and the mridangam are recorded at pitch C Sharp. The pitches of singers 2 and 3 are D and G respectively. So there needs to be a shift in the pitch of the recorded violin and mridangam. The pitch shifting is done using the librosa library in Python for both singers [35]. For Inter Augmentation, all the tracks obtained post Intra augmentation, are chosen to be augmented. As mentioned in Section 3.3.1, the stems are recorded at 60 BPM and most Carnatic music recordings spanned between 60BPM-90BPM. There seemed a possibility for time stretching the data to have instances of 75BPM and 90BPM data. The existing data is time-stretched using the librosa library in python [35] to have the above instances of data.

The creative freedom taken as part of the augmentation are informed choices. Having artificial data as part of the dataset would increase the instances of data, yet induce biases. In spite of that, several researchers have used data augmentation techniques to improve MIR tasks such as classification [36, 37]. The decisions taken are based

on the tradition. The choice to have 75 and 90 BPM instances is based on an analysis that Carnatic Music recordings are rendered between speeds of 60-90 BPM. Pitch shifting the violin and mridangam induces timbral changes in the augmented data. A manual listening of the pitch-shifted data is done to understand if the credibility is below threshold and then the data is added to the Train set. Data with extremely high changes in timbral quality is added to the Testing set.

3.4.1 The Mixing Process

The Mixing process is completely done on GarageBand. The process has two steps; The first step includes manually inspecting the stems and checking if they align with each other. There are several instances where a particular melody was sung twice and it required manual correction through thorough listening. The stems are time-aligned because Source Separation algorithms need time-aligned stems as input for training.

The second step is Levelling and Compression. Given that the main melody is the vocals and the other stems are recorded in a similar setting, the violin had to be lowered accordingly. The mridangam and the tanpura stems are added and mixed by ear using the reference as the Saraga dataset. This step is crucial because having a mix close to that of the datasets in the literature can help better generalization.

The third step is mixing the Accompaniment stems. Since we are aiming to use a two-stem model, the mridangam, tanpura and violin stems are mixed to have an Accompaniment stem and then the Vocal stem is mixed with the Accompaniment stem to have the Mix file. All the stems are stereophonic, of the WAV format and encoded at 44.1 KHz.

3.4.2 Challenges and Creative choices in the process

There were quite a few challenges while working on building the dataset. A major challenge was remotely coordinating and recording artists. Carnatic music is extremely region-specific and most practitioners live in the Southern States of India.

With no previous experience recording music remotely, this was a huge challenge to overcome. Coordinating with artists and the loop of correcting mistakes and re-recording is a time-consuming process. When working over a set time frame, this process proves to be challenging.

Another challenge was the unexpected shortcomings due to the lack of equipment. This is very related to the lack of remote recording experience. These challenges led to more experimentation and resulted in improvement with time. These challenges also lead to having more instances of data of different types which could possibly have a better impact on the inference.

Wrapping up the discussion, this section elaborated on the process of building the dataset, the augmentation techniques used and the creative choices behind the process.

Chapter 4

The Retraining and Fine-tuning Process

This Chapter elaborates on the retraining and fine-tuning process to help build custom Source Separation models for the use case of Carnatic music. In section 4.1, the model Spleeter is defined and the choice of model is justified. The architecture of the model is briefly discussed. In section 4.2, the prerequisites for the retraining process are listed and the hyperparameter configurations are discussed in detail. The next section demonstrates the Model retraining and finetuning process. Section 4.3.3 discusses the Loading and Validation of the model. This also includes how to use the custom models and run inference with them.

This section is a detailed description of the retraining process of the Spleeter Source separation model. The audio data is also processed according to the model requirements. Note that different models can have different data-related requirements.

4.1 The Model and Prerequisites for the Retraining process

Out of the models listed in Section 2.3, Spleeter is chosen for our task. Given that we have a medium-sized dataset, using a spectrogram-based model like spleeter might

help achieve better results. The other reason is that Spleeter is widely used for MIR research in Indian Art Music and Demucs struggles to cope with the characteristics of the tradition as mentioned in Section 2.3.3.

The Prerequisites can be divided into two; The first is data-related prerequisites to run the model. The other is configuration files containing details relevant to the dataset and the model. The details pertinent to the model include the hyperparameter configurations and the model checkpoint details.

For Spleeter, audio data needs to be encoded at 44.1 KHz. All stems of all tracks must be stereophonic. The data should preferably be of the mp3 format. All stems need to be time aligned and the mix must be the sum of the 2 stems- Vocals and Accompaniment. There are no size-related constraints on the data used for training. The dataset should be divided into three sets; Training, Validation and Test. The training and validation sets are used for training and fine-tuning the hyperparameters respectively. The testing is used to run inference and evaluate the results.

The configuration files include two comma-separated value (CSV) files for the train and validation sets. This file should contain the stem paths of the respective tracks- Mix, Vocals and Accompaniment to the Train and Validation files in three separate columns. These files should also contain the duration of each track in seconds. It is assumed that all the stems of each track have the same duration due to the requirement that they are time-aligned. The other files include a JavaScript Object Notation (JSON) file that contains details of the model that is used and the data that we are retraining it with.

The data-related details include the path to the train and validation CSV files, the sample rate of the audio files, the number of channels to denote whether the files are monophonic or stereophonic, and the frame length and hop size. The model-related details include the number of stems of the model (excluding the mix), the description of the stems, the name of the model, the train and validation cache directories, the maximum training steps, the number of checkpoint steps, the batch size, and the learning rate.

The data-related details are already part of the example JSON file that is used as a reference. The data preparation is done according to the reference file to maintain uniformity. The maximum training steps is set to 2000000 and the learning rate is set to 1e-4. We should note that these model-related parameters are general values that are specified by Spleeter. In the later sections, we will see how experimenting with these parameters can help the model learn better.

4.2 Model Retraining and fine-tuning process

With the prerequisites ready, we start the model training. The model retraining is done with the 2 stem model offered by Spleeter. Note that the retraining process can be computationally intensive, so training on a computer with a Graphics Processing Unit (GPU) can help complete the training process faster. In our case, the training is done over a computer powered by Nvidia V100 Tensor core GPUs. The model retraining is a multi-step process. It starts with cloning the Spleeter repository to our desired directory in the computer. The Spleeter repository contains sub-folders containing the existing pre-trained models and their weights, the configuration files and others. To this directory, move the folder containing the Carnatic Source Separation Dataset that we have curated. The dataset folder should have sub-folders containing the training and the testing set. Note that the validation set is a subset of the training set.

To train the model with Carnatic music datasets, the model checkpoint is set to 200000 and the learning rate is set to 1e-5 (as opposed to 1e-4 suggested by spleeter). The learning rate is set to a lower number to particularize better for the use case [38]. The configuration files that include the CSV files describing the train and the validation set respectively, and the JSON file describing the model are also added to the spleeter folder. With these in place, the training process can be initiated by writing a command line code given by spleeter.

The code is: *spleeter train -p config.json -d /path/to/dataset*

Here, config.json denotes the configuration file with the information of the model,

and the path/to/dataset is the directory of the dataset(in this case the Carnatic music source separation dataset) that we are aiming to retrain the model with.

The fine-tuning process is very similar to the retraining process with the difference being that the model is fine-tuned over the existing pre-trained 2-stem model. The retraining process is when the model is trained from scratch and needs large amounts of data to learn. Given that the pre-trained 2-stem spleeter model is used for the Source Separation of Carnatic music data, fine-tuning with Carnatic music data can have a greater impact on the separations. For the fine-tuning process, a list of models is made available by spleeter to continue training. The 2-stem finetune model is downloaded and explicitly indicated as the model name in the JSON file. The maximum training steps are set to a number higher than 1000000 because this number indicates the training steps of the pre-trained model. The rest of the process remains the same.

4.3 Retraining and Fine-tuning Experiments

There are two models built as part of the retraining and fine-tuning process. Each of the models, data and training process are elaborated below. For the first instance of retraining, the 2 stem model is retrained with the dataset we have built. The model checkpoints are loaded for every 300 steps. The retraining process takes approximately 42 hours to complete 2000000 steps. The generated model is saved in the pre-trained models folder in the spleeter directory. This model is named as Model 1. The training loss is obtained as 0.154 with an accuracy of 84.6 per cent. The intention of training the model with the data is to familiarize the model with clean Carnatic music stems

For the second instance, the fine-tuning route is chosen. 100 multitracks of Saraga are mixed in a similar fashion as recorded in Section 3.4.1 and added to the existing dataset. The model is fine-tuned with the updated dataset with Saraga stems. The model is retrained with the same parameters as Model 1. As stated in Chapter 2, the state-of-the-art datasets are directly derived from concert recordings. The intention

behind training with Saraga stems with bleeding is to generalize better for existing datasets in the literature. 100 multitracks are chosen from the Saraga dataset to have a good ratio of reference stems that we aim to generalize for. This model is named as Model 2. The training loss is obtained as 0.149 with an accuracy of 85.1 per cent.

The finetuned models are pre-trained models that allow training with a different dataset to improve the separation, training these models can provide improved results for the tradition. Model 2 is generated in an effort to generalize well for datasets that we have as part of the Dunya corpora [39]. Given that the multitrack data in Saraga Carnatic has severe bleeding, using just data with bleeding cannot help obtain clean vocal stems. The dataset that we have created can possibly help the model learn with clean stems and data with bleeding can help generalize better for the type of datasets that we aim to work with as part of Carnatic-MIR.

4.3.1 Loading and Validating the model

Post the model training or fine-tuning models, these models are moved to the pre-trained model's subfolder within the spleeter folder. While separating using spleeter, the default folder where spleeter looks for the models should be set to the pre-trained model's folder. This could be done by setting the environment path as the directory to this folder, within the bashrc file of the system. If not set to this directory, spleeter will not be able to locate the model and instead download the pre-trained 2-stem model to facilitate separation.

To validate the model and check if the separations take place, the spleeter separate command is extremely useful. To run separations, the following code is run from the spleeter folder directory

spleeter separate audio-example.mp3 -o audio-output -p config.json

In this code, audio-example.mp3 is the audio sample that you want to separate, audio-output is the folder in which the separated stems will be saved, and config.json refers to the configuration file containing the details of the model and the data to be

loaded.If the model is loaded and used, the two stems- Vocal and Accompaniment will be saved as two separate audio files. A manual listening of the separated stems would help identify if the stems are separated or not.

4.4 Conclusions

The model retraining and fine-tuning processes are elaborated in detail. The Spleeter model is chosen to retrain and fine-tune due to its frequent usage in Carnatic-MIR research and also for its range of adaptability. Section 4.1 lists the prerequisites to start with the retraining and fine-tuning process and Section 4.2 details the process itself. Section 4.3 elaborates on the models built by experimenting with two sets of data: The Carnatic Multi-track dataset and, the Saraga multi-track dataset built specifically for the fine-tuning process. Section 4.3.1 briefs about loading the models built with custom Carnatic music data and validation of the model.

Chapter 5

Evaluation and Results

In this section, we evaluate the models that are explained in section 4.2.2. Section 5.1 describes the Source to Distortion Ratio metric that is generally used for benchmarking and evaluating the inference. Python libraries like museval are also used to evaluate Source separation algorithms with the reference dataset as MUSDB18 [40]. A manual evaluation approach is taken and the Source to Distortion Ratio is calculated for all tracks. Section 5.2 discusses the results generated using Model 1 against the baseline Spleeter 2-stem model. A perceptual evaluation of how the stems separated by the two models differ is elaborated on. Section 5.3 discusses the results generated on the testing set using Model 2. A perceptual evaluation of the model over the Saraga dataset is undertaken due to the lack of clean ground truth data.

The evaluation is based on energy-related metrics like the SDR and also through perception-based metrics and manual listening [41]. Given that Source Separation is eventually used for tasks like melodic analysis, a critical analysis of the Source separated stems in terms of bleeding is desirable. The last section compares the vocal stems separated by the three models(Model 1, Model 2 and baseline Spleeter) by analysing the bleed from different instruments.

5.1 The Source to Distortion Ratio

An estimate of a source constitutes four separate components; Target source, interference, noise, and artifacts. The Source to Distortion Ratio is a metric to evaluate how good a source sounds against the ground truth reference signal. To calculate the Source to Distortion Ratio, the reference ground truth signal array of the Vocals and Accompaniment, and the array of stems obtained through source separation are required. The Source to Distortion Ratio is calculated over the Test set with both the models against baseline spleeter. It is calculated as follows:

$$\mathrm{SDR} := 10 \log_{10} \left(\frac{\| s_{\mathrm{target}} \|^2}{\| e_{\mathrm{interf}} + e_{\mathrm{noise}} + e_{\mathrm{artif}} \|^2} \right)$$

Figure 4: Calculation of the Source to Distortion Ratio

The Source to Distortion Ratio is the most commonly used metric to evaluate Source Separation, but studies suggest that a perceptual test could prove to be more beneficial as it is the truest reflection of the human sense. Perceptual tests can differ subjectively, therefore these tests should be conducted over a large set of people. This process can be extremely time-intensive and laborious.

5.2 Results from Model 1

Model 1 here refers to the model that was retrained from scratch using the remotely recorded Carnatic Source Separation dataset. The Test set has about 12 tracks. The mixes of each of the 12 tracks are separated using Model 1 and the baseline spleeter. The separated stems are compared against the ground truth and the Source to Distortion Ratio is calculated for both stems of each track. The median SDR values are presented in Table 2 for discussion.

Baseline Spleeter gives SDR values in the range of 6-7 dB when evaluated over the

Table 3: Median SDR values

Model	**Vocal SDR**	**Accompaniment SDR**
Baseline Spleeter	5.3205	5.2204
Model 1	6.40055	6.375

MUSDB18 dataset and for the use case of Carnatic music it does not perform up to that level. From the above table, the SDR value offered by Baseline Spleeter is 5.3205 for the Vocal stem. When the kind of bleed is analysed, there is a strong presence of the attacks of the mridangam. In spite of Spleeter being trained with Drum stems, it is not able to separate the onset of mridangam completely. This could be because of their lack of familiarity with Spleeter with pitched percussion instruments. The other type of bleed is the overlap of the violin.

Model 1 gives an SDR value of 6.40055 dB when evaluated over the test set. It outperforms baseline spleeter by a margin of 1 dB. Perceptually, baseline trained spleeter manages to remove the mridangam attacks and overlapping violin considerably. Model 1 consistently outperforms baseline spleeter for every single track in the test set. As mentioned in Section 3.3, the dataset with which Model 1 is trained does not mirror the recording quality of the datasets in the literature. To test the efficiency of the model, separations are run over the Saraga dataset. Due to the lack of ground truth data, a perceptual evaluation of the vocal stem separated by both models is done.

Model 1 was not able to generalize well for the Saraga dataset. This could be due to the fact that the model was trained using a low learning rate which helped it particularize for the Carnatic multi-track dataset set, but not for the Saraga dataset. Given that this dataset helped cope with bleed-related issues that Spleeter could not mitigate, it could be assumed that fine-tuning with bleed-prone stems from the Saraga multitrack plus the Carnatic multi-track dataset could help the model overcome these issues.

5.3 Results from Model 2

Model 2 is used to separate the stems over three tracks from the Saraga Carnatic dataset; Ananda Sagara, Chenduril Nindradum and Bhavamulona. Baseline Spleeter is also used to separate the stems over these tracks. In spite of the fact that these tracks have multi-track data, they cannot be used as ground truth because the multi-track data have severe leakage of the other instruments. Hence, a metric-based evaluation cannot be made over these tracks. A small comparison of the separated tracks is made by manually listening to the tracks.

Baseline spleeter continues to struggle with issues pertaining to bleeding. Model 2 manages to almost completely remove the bleeding related to the percussion in spite of being trained majorly with data that has severe bleeding. The violin overlap-related bleeding is also considerably low. The fact that there are tracks with clean stems through the dataset we recorded could account for this behaviour.

Model 2 is also used to separate stems from the test set containing tracks from the above-created dataset. The model underperforms in this case in spite of being fine-tuned with both datasets. Considering the fact that there is a higher ratio of Saraga tracks, the model could have overfit for this type of data. This could be the reason that the model performs extremely well in the case of Saraga and underperforms with the Carnatic multi-track Dataset.

The results prove to be promising, yet need a thorough inspection through a perceptual test conducted with either musicians or researchers that have previously worked with Source separation algorithms. Implementing a perceptual listening test like MUSHRA [42] could be a possible way to better evaluate the Source separated stems.

Chapter 6

Conclusions and Future Work

In this section, we summarize and derive conclusions from the work done as part of the thesis. A section is dedicated to discussing the work that could be done on top of this thesis and the possible future directions of research.

6.1 Summary of the thesis

The thesis is a combination of musicologically informed knowledge and Deep Learning based technologies. Chapter 1 of the thesis is an introduction to the fundamentals of Music Information retrieval, Carnatic Music, and Source Separation and poses a key question. Why is Source Separation a relevant problem for research in Carnatic music? This helps us define the objectives of the thesis including the creation of a multitrack dataset with Carnatic music stems, reusing existing multitrack recordings from datasets part of the literature, and retraining or finetuning with existing state-of-the-art models used for Source separation. Chapter 2 is a gentle introduction to MIR research in Indian art music, state-of-the-art technologies used for Source Separation, datasets that have been used to train these models and the music demixing challenge.

Chapter 3 describes the creation of a Carnatic music-specific multitrack dataset in great detail. The biases in existing multitrack datasets are discussed and the stems

in Carnatic music are listed. The choice of stems for the dataset is justified and the list of compositions is discussed. The complete recording process and the possible shortcomings are elaborated. The data augmentation techniques used to increase the size of the dataset are discussed and concerns related to the credibility of the data are raised and justified. Finally, the challenges and the creative liberties are listed, questioned and justified.

Chapter 4 introduces the retraining and fine-tuning process. The model-Spleeter is defined and the prerequisites for retraining are listed. The model retraining process is introduced and the differences between retraining and fine-tuning are explained. The models built as part of the experiments are listed and the data choices are justified. Finally, the steps for loading and validation of the model are listed.

Chapter 5 introduces the evaluation metric called Source to Distortion Ratio and discusses the need for perceptual evaluation of the source-separated stems. The results obtained through Model 1 are evaluated using metrics, and also through manual listening. A manual listening of the stems separated using Model 1, Model 2 and pre-trained spleeter is done and a perceptual comparison is made in terms of the bleeding of the accompanying instruments. In an effort to generalize for the datasets in the literature pertaining to Carnatic music, Model 2 is fine-tuned with data with leakage from the Saraga dataset in addition to the dataset used to train Model 1. The results are analysed and conclusions are extracted.

6.2 Extracted Conclusions

The motivation of the thesis was to improve source separation for the case of Carnatic music. Through Chapter 5, the following conclusions are derived:

1) Model 1 performs better than baseline spleeter on the data it is trained on. The model manages to considerably mitigate mridangam onsets and the overlapping violin to. It could not generalize for the case of the Saraga dataset due to the fact that Saraga has a completely different recording quality.

2) Model 2 manages to outperform baseline spleeter in terms of perceptual quality

when evaluated over the Saraga dataset. In spite of being trained predominantly with data with leakage, the model manages to almost eliminate percussion-related attacks and overlap of the violin, generalizing well for the Saraga dataset. It could be assumed that the dataset used to train Model 1 had large amounts of clean stems accounting for the model managed to learn better.

3) Model 2 underperforms over the test set of the Carnatic Source Separation dataset in spite of being trained with tracks of the dataset. The possible reason could be that the model overfit for the Saraga dataset due to the fact that it is trained with more tracks from Saraga than the Carnatic Source Separation dataset.

4) A thorough perceptual study is required to evaluate Model 2 over the Saraga dataset due to the lack of ground truth data.

6.3 Possible research directions

The thesis completely focuses on improving Source separation for the use case of Carnatic music. The approach used for the thesis is completely data-driven and most choices that are informed through tradition have implications on the way the data is built and used. There are no implications on the model and the way the model could be tweaked to improve the task at hand. A more tradition-informed model-based approach would possibly help improve Source separation to a larger extent. The other approach that could be taken is concentrating on pre-processing or post-processing steps that could help combat the major issue of leakage. These steps could include studying the types of leakage and building a leakage removal pipeline specific to Carnatic music.

A completely different direction will be using different modes of data to improve source separation. Given that Carnatic music relies on impromptu responses from accompanying musicians and using standardized data cannot help generalise for all recordings, working with the video modality and understanding how motion information can be captured can be extremely beneficial. This multi-modal approach can be an effective way to mitigate issues that current source separation models

struggle with.

A study of how improving Source Separation can impact the melodic analysis could be a possible direction that could be taken. The intention of the thesis was to improve Source Separation for Carnatic music, but the motivation behind was to improve the melodic analysis. A comprehensive study of how the improvement in Source Separation impacts tasks like motif identification and raga classification can be useful for an overall understanding of the task.

Finally, the thesis is based on improving Source Separation for the use case of Carnatic Music. Hindustani Music is another classical music form that is a cousin of Carnatic music, that is researched in the MIR community. Hindustani music is a tradition that has several similarities and differences with Carnatic music. There are similar issues with Source separation the art form faces when used for obtaining individual stems [43, 44]. Given the context, analysing if the models retrained and fine-tuned with Carnatic music can improve Source separation for Hindustani music can help understand the generalizability of the model for the combined case of Indian Art Music.

List of Figures

1 The Process of Music Source Separation 3

2 The architecture of a U-Net 13

3 A Carnatic Music Concert 17

4 Calculation of the Source to Distortion Ratio 33

List of Tables

1 List of Compositions with its respective raga and taala 21

2 List of Compositions with singer-related information and duration . . 23

3 Median SDR values 34

Bibliography

[1] Casey, M. A. *et al.* Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE* **96**, 668–696 (2008).

[2] Serra, X. A multicultural approach in music information research. In *Klapuri A, Leider C, editors. ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference; 2011 October 24-28; Miami, Florida (USA). Miami: University of Miami; 2011.* (International Society for Music Information Retrieval (ISMIR), 2011).

[3] Krishna, T. & Ishwar, V. Carnatic music: Svara, gamaka, motif and raga identity. In *Serra X, Rao P, Murthy H, Bozkurt B, editors. Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12-13; Istanbul, Turkey. Barcelona: Universitat Pompeu Fabra; 2012.* (Universitat Pompeu Fabra, 2012).

[4] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I. & Bittner, R. Musdb18-a corpus for music separation (2017).

[5] Bittner, R. M. *et al.* Medleydb: A multitrack dataset for annotation-intensive mir research. In *ISMIR*, vol. 14, 155–160 (2014).

[6] Hennequin, R., Khlif, A., Voituret, F. & Moussallam, M. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software* **5**, 2154 (2020).

[7] Défossez, A., Usunier, N., Bottou, L. & Bach, F. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174* (2019).

[8] Stöter, F.-R., Uhlich, S., Liutkus, A. & Mitsufuji, Y. Open-unmix-a reference implementation for music source separation. *Journal of Open Source Software* **4**, 1667 (2019).

[9] Sebastian, J. & Murthy, H. A. Group delay based music source separation using deep recurrent neural networks. In *2016 International Conference on Signal Processing and Communications (SPCOM)*, 1–5 (IEEE, 2016).

[10] Srinivasamurthy, A., Gulati, S., Repetto, R. C. & Serra, X. Saraga: Open datasets for research on indian art music. *Empirical Musicology Review* **16**, 85–98 (2021).

[11] Le Roux, J., Wisdom, S., Erdogan, H. & Hershey, J. R. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 626–630 (IEEE, 2019).

[12] Serra, X. Creating research corpora for the computational study of music: the case of the compmusic project. In *Audio engineering society conference: 53rd international conference: Semantic audio* (Audio Engineering Society, 2014).

[13] Srinivasamurthy, A., Koduri, G. K., Gulati, S., Ishwar, V. & Serra, X. Corpora for music information research in indian art music. In *Georgaki A, Kouroupetroglou G, eds. Proceedings of the 2014 International Computer Music Conference, ICMC/SMC; 2014 Sept 14-20; Athens, Greece.[Michigan]: Michigan Publishing; 2014.* (Michigan Publishing, 2014).

[14] Koduri, G. K., Ishwar, V., Serrà, J. & Serra, X. Intonation analysis of rāgas in carnatic music. *Journal of New Music Research* **43**, 72–93 (2014).

[15] Srinivasamurthy, A., Holzapfel, A. & Serra, X. In search of automatic rhythm analysis methods for turkish and indian art music. *Journal of New Music Research* **43**, 94–114 (2014).

[16] Srinivasamurthy, A. & Serra, X. A supervised approach to hierarchical metrical cycle tracking from audio music recordings. In *2014 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), 5217–5221 (IEEE, 2014).

[17] Gulati, S., Salamon, J. & Serra, X. A two-stage approach for tonic identification in indian art music. In *Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12-13; Istanbul, Turkey. Barcelona: Universitat Pompeu Fabra; 2012. p. 119-127* (Universitat Pompeu Fabra, 2012).

[18] Gulati, S. *et al.* Automatic tonic identification in indian art music: approaches and evaluation. *Journal of New Music Research* **43**, 53–71 (2014).

[19] Salamon, J., Gulati, S. & Serra, X. A multipitch approach to tonic identification in indian classical music. In *Gouyon F, Herrera P, Martins LG, Müller M. ISMIR 2012: Proceedings of the 13th International Society for Music Information Retrieval Conference; 2012 Oct 8-12; Porto, Portugal. Porto: FEUP Ediçoes; 2012.* (International Society for Music Information Retrieval (ISMIR), 2012).

[20] Défossez, A., Usunier, N., Bottou, L. & Bach, F. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254* (2019).

[21] Kaspersen, E. T., Kounalakis, T. & Erkut, C. Hydranet: A real-time waveform separation network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4327–4331 (IEEE, 2020).

[22] Markovic, D., Defossez, A. & Richard, A. Implicit neural spatial filtering for multichannel source separation in the waveform domain. *arXiv preprint arXiv:2206.15423* (2022).

[23] Stoller, D., Ewert, S. & Dixon, S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185* (2018).

[24] Oh, J., Kim, D. & Yun, S.-Y. Spectrogram-channels u-net: a source separation model viewing each channel as the spectrogram of each source. *arXiv preprint arXiv:1810.11520* (2018).

[25] Défossez, A. Hybrid spectrogram and waveform source separation. *arXiv preprint arXiv:2111.03600* (2021).

[26] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241 (Springer, 2015).

[27] John, S., Sinith, M., Sudheesh, R. & Lalu, P. Classification of indian classical carnatic music based on raga using deep learning. In *2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 110–113 (IEEE, 2020).

[28] Dawalatabad, N. *et al.* Front-end diarization for percussion separation in tani-avartanam of carnatic music concerts. *arXiv preprint arXiv:2103.03215* (2021).

[29] Nuttall, T., Plaja-Roglans, G., Pearson, L. & Serra, X. The matrix profile for motif discovery in audio-an example application in carnatic music. In *International Symposium on Computer Music Multidisciplinary Research*, 228–237 (Springer, 2021).

[30] Mitsufuji, Y. *et al.* Music demixing challenge 2021. *Frontiers in Signal Processing* **1**, 808395 (2022).

[31] Holzapfel, A., Sturm, B. & Coeckelbergh, M. Ethical dimensions of music information retrieval technology. *Transactions of the International Society for Music Information Retrieval* **1**, 44–55 (2018).

[32] Mani, C. Customised pedagogical tools to aid aural-oral transmission: Raga-curve and gesture. *The Finnish Journal of Music Education* **21**, 39 (2018).

[33] Morris, R. Variation and process in south indian music: Some kritis and their sangatis. *Music Theory Spectrum* **23**, 74–89 (2001).

[34] Ramires, A. & Serra, X. Data augmentation for instrument classification robust to audio effects. *arXiv preprint arXiv:1907.08520* (2019).

[35] McFee, B. *et al.* librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, vol. 8, 18–25 (2015).

[36] Nanni, L., Maguolo, G. & Paci, M. Data augmentation approaches for improving animal audio classification. *Ecological Informatics* **57**, 101084 (2020).

[37] Aguiar, R. L., Costa, Y. M. & Silla, C. N. Exploring data augmentation to improve music genre classification with convnets. In *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE, 2018).

[38] Smith, L. N. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820* (2018).

[39] Porter, A., Sordo, M. & Serra, X. Dunya: A system for browsing audio music collections exploiting cultural context. In *Britto A, Gouyon F, Dixon S. 14th International Society for Music Information Retrieval Conference (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 101-6.* (International Society for Music Information Retrieval (ISMIR), 2013).

[40] Stöter, F.-R., Liutkus, A. & Ito, N. The 2018 signal separation evaluation campaign. In *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Surrey, UK*, 293–305 (2018).

[41] Emiya, V., Vincent, E., Harlander, N. & Hohmann, V. The peass toolkit-perceptual evaluation methods for audio source separation. In *9th Int. Conf. on Latent Variable Analysis and Signal Separation* (2010).

[42] Series, B. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly* (2014).

[43] Clayton, M., Rao, P., Shikarpur, N., Roychowdhury, S. & Li, J. Raga classification from vocal performances using multimodal analysis. In *Ismir 2022 Hybrid Conference* (2022).

[44] Shikarpur, N. N., Keskar, A. & Rao, P. Computational analysis of melodic mode switching in raga performance. In *ISMIR*, 657–664 (2021).