Experiments in Automatic Sample Detection in Hip Hop Music

Trey Bradley

Submitted in partial fulfillment of the requirements for the
Master of Music in Music Technology
in the Department of Music and Performing Arts Professions
Steinhardt School
New York University

Advisor: Dr. Brian McFee
Reader: Dr. Robert Rowe

February 19, 2023

Abstract

Sampling in this context dates back to mid-century developments in music technology and composition that dealt with recording environmental sounds and embedding the recording into new compositions. As recording and production technologies developed, sampling reached a larger group of producers and composers, predominantly occupying the Hip-Hop and Electronic music genres. With its large-scale impact on music production and consumption, sampling has received attention from fans, musicologists and legal entities who ask questions about musical influence, the evolution of composition, musical plagiarism, and cultural policy. However, a system to determine cases of sampling with high accuracy that accounts for new cases and styles of sampling that are released onto the web everyday does not yet exist. This thesis project draws on prior work to evaluate automatic sample detection experiments. Previous attempts at this task included audio fingerprinting and non-negative matrix factorization (NMF) approaches. Given its promise, this project experimented with the NMF-based approaches, which frame sample detection as a source separation task. A ground truth database was compiled and experiments were implemented to evaluate the different systems' abilities to detect samples in query songs. Ultimately, this project offered immediate takeaways and insight into the use of various NMF implementations to successfully detect samples in query songs. The strongest system obtained 73% accuracy. In the end, the problem is not considered completely solved, as future work could be completed to increase the accuracy and speed of the systems.

Acknowledgements

I would like to thank my mother and family for supporting me in my music technology pursuits and for providing me with a conducive environment to work on this thesis project in a healthy way, amidst covid-19.

Thank you to my thesis advisor for guidance along the way. Dr. McFee's suggestions and support was extremely helpful to this project and my music technology education.

Finally, I would like to acknowledge previous authors of automatic sample detection projects who wrote about this topic and engaged in correspondence regarding their implementations.

Contents

1 Introduction 7

1.1 Goals and Motivations . . . 7

1.1.1 A Music Discovery Tool . . . 8

1.1.2 A Cultural Policy Tool . . . 9

1.2 Evolution of Sampling . . . 9

1.2.1 Musique Concrète . . . 10

1.2.2 Electronic Music in Germany and Japan . . . 11

1.2.3 Turntablism and Sampling in Hip-Hop . . . 12

1.3 Early Sampling Technology . . . 14

1.4 Content-Based Audio Analysis . . . 16

1.4.1 Time/Frequency Representations of Audio . . . 17

1.4.2 Harmonic-Percussive Source Separation . . . 19

1.4.3 Non-Negative Matrix Factorization (NMF) . . . 20

1.4.4 Musically-Informed NMF . . . 22

1.4.5 Aligning Musical Sequences . . . 23

1.5 Related Work . . . 25

1.5.1 Audio Identification . . . 25

1.5.2 Cover Song Detection . . . 26

2 Prior Work in Automatic Sample Detection 27

2.1 Landmark-Based Audio Fingerprinting . . . 27

2.2 Proposing a System to Detect Music Plagiarism . . . 29

2.2.1 Implementing PFNMF in Automatic Sample Detection . . 31

2.2.2 PFNMF with DTW Similarity Measure . . . 32

2.2.3 Selecting Query-Sample Candidates . . . 34

3 Methodology and Results 37

3.1 Ground Truth Data . . . 37

3.2 Evaluation Metrics . . . 39

3.3 Baseline Experiment . . . 39

3.3.1 Results and Discussion of the Baseline System . . . 41

3.3.2 Results and Discussion of a Second Baseline System . . . 42

3.4 Experiment 1: Fully Fixed NMF . . . 44

3.4.1 Results and Discussion of Experiment 1: Fully Fixed NMF 44

3.5 Exploring Dictionary Information and NMF Initialization . . . 45

3.6 Experiment 2: Traditional NMF . . . 46

3.6.1 Results of Experiment 2: Traditional NMF 46
3.7 Experiment 3: NMF with Custom Initialization 47
3.7.1 Results of Experiment 3: NMF with Custom Initialization 48
3.8 Experiment 4: Isolating Query Bases 48
3.8.1 Results of Experiment 4: Isolating Query Bases 49
3.9 Experiment 5: Differentiating Harmonic and Percussive Sample . 50
3.9.1 Results of Experiment 5: Differentiating Harmonic and Percussive Sample 51
3.10 Discussion 51

4 Conclusion 53

Bibliography 55

List of Figures

1 E-Mu SP12 from Harkins (2019) 15

2 Time-domain representation of two songs, where a sample relationship occurs . 17

3 Magnitude spectrograms of two songs, where a sample relationship occurs . . 17

4 Hpss of a complex audio mixture 20

5 NMF decomposition on a magnitude spectrogram 21

6 Matching Function of a sub-sequence alignment of a sample that appears 4 times in the query 24

7 A spectrogram and its constellation map. Adapted from Wang (2003) 25

8 Block Diagram of Proposed NMF System 30

1 Introduction

In the decades leading up to the turn of the 20th century, new concepts of time and the connectivity of the universe were developing in scientific communities around the world. Artists and philosophers at the time explored the synchronicity and synthesis in time and space between objects and ideas that were not conventionally related. For instance a German word, Gesamtkunstwerk, was written about in 1827 by philosopher K.F.E. Trahndorff, referring to an ideal synthesis of the arts that comprised of multiple art forms, disciplines, and media. Richard Wagner later applied the idea of Gesamtkunstwerk to combine several artistic media in his works, followed by other composers, like Igor Stravinsky, who combined rhythmic, melodic, and harmonic content from a variety of styles and cultures (Burkholder et al., 2014). By the middle of the 20th century, composers like John Cage and Pierre Schaeffer began experimenting with sound recording technology to combine sounds from locomotives, car horns, and other environmental noises in their musical works. As recording and production technologies advanced, the removal of contexts and embedding of audio segments in music, which came to be established as sampling, grew. Hip-Hop and Electronic music producers adapted this technique as an integral technique to their styles. This thesis project explores sampling, the reuse of previously recorded sounds in new audio mixtures, from a Music Information Retrieval (MIR) approach.

Early music technologists who experimented with sound framed music as a sum of its feature-bearing components. With regard to detecting samples in new recordings, the MIR approach of this thesis frames sound in a similar way. After establishing the goals, motivations, and requirements of a state-of-the-art sample detection system, the history of sampling and background information on content-based audio analysis is provided in Section 1. Next, prior MIR attempts at automatic sample detection are reviewed in Section 2. Section 3 discusses the methodology and results. Finally, section 4 analyzes the results of this project and discuss its impact for future work.

1.1 Goals and Motivations

This project's automatic and content-based approach to sample detection uses concepts from MIR and machine learning (ML) with the goal of creating an objective algorithm that can irrefutably determine the existence of prerecorded musical audio, a sample, within a new recording. The task is difficult because the new recording, referred to in this paper as a query track, combines the sample with a mixture of other audio sources. Therefore, the system needs to be able to distinguish between the audio of the sample and the audio of the remaining mixture, before identifying the sample. Furthermore, the length of the sample and time location within the query track is unknown to the system. Lastly, the task is made more difficult since production and composition tools give sampling artists creative control to process and manipulate the sample beyond recognition in the query track. The manipulations that a sample can undergo include pitch shifting, time stretching, reverse, changes in level, and a host of other distortions and signal processes. While computers can be trained to recognize patterns in audio despite these manipulations to its content, designing such a computational system is not a trivial task.

1.1.1 A Music Discovery Tool

Sampling requires that a composer or producer undergoes a search to obtaining the samples. These artists have an intended use for samples and actively confront their search with intent and preference. After undergoing unique processes to discovering a sample, the material often becomes integral to the artistic expression, message and musical vision of the composer. Studying sampling culture and the creative choices of these artists can highlight patterns of musical preferences, styles, ideas, and sources of inspiration and influence. Samples can also reflect cultural themes and values of the artist and its audience. For example, on the 1988 release, It Takes a Nation of Millions to Hold Us Back, Public Enemy and their producers consciously included politically-oriented audio samples and delivered the album "front-loaded with sirens, squeals, and squawks that augmented the chaotic backing tracks over which front man Chuck D laid his politically and poetically radical rhymes" (McLeod and DiCola, 2011, p.22). This record exemplifies the creative use of carefully researched and produced samples to shape the ethos of a work, which reflected the social, political and cultural climate regarding the state of African American civil rights in the United States during the 1990s. In a broader context, the deliberately chosen and placed samples can expose a wealth of information about the days, times, and experiences of artists, their audiences, and the societies they inhabit. This project can help provide an efficient and accurate research and discovery tool for fans and musicologists who desire knowledge about the vast networks of influence between artists, genres, and musical ideas, highlighting their transcendence through time and cultures.

WhoSampled.com, is a resource that collects data about sampling and networks of influence between musicians. While its data provides useful information and thousands of specific sample relations between query and sample songs, it also exposes fans and researchers to the limitations of human-error from crowd-sourced data. In searching through thousands of examples in their database, some of the examples misrepresented lyrical or melodic interpolations as sample relations. This further motivated an automatic and irrefutable detection method that relied on a standard definition of sampling and audio analysis to verify sampling claims. With today's unprecedented amounts of audio being delivered to the web, an automatic approach to determining potential cases of sampling would be needed to process the vastness of that information in a way that removed human error and inefficiencies.

1.1.2 A Cultural Policy Tool

While this project is mainly motivated to serve musicologists, researchers, and fans in their exploration of music and culture, legal concerns around creativity, originality and musical plagiarism also arise when dealing with matters of intellectual property (IP). An example of a cultural policy is the IP (copyright) clause in the U.S. Constitution which imposes constraints on the production of art and culture in the U.S., granting the federal government the right to define and govern "the progress of science and useful arts" (U.S. Const. art. I, §3, cl. 8). As the 1980s saw an increase in sampling and the commercial successes of Hip-Hop and Electronic music, legal systems and owners of IP around the world took notice and began applying the IP clause in claims of musical citation and plagiarism. Today, discussions about how to govern digital sampling persist, as digital technologies increase the transmission of IP, accelerating with the digital boom and large social media platforms that host user-generated content, which commonly features copyrighted audio samples. Since automatic sample detection is a content-based analysis method, legal entities and institutions can use this approach to help establish standards of fair use, since one of the factors that establishes fair use is the length of a sample and its perceivable presence in the derivative work. Ultimately this project does not dissect cultural policy or the governance of creativity in media and art, but it has the potential to offer accurate data where cases of copyright infringement occur and can help manage ownership and disputes over IP.

1.2 Evolution of Sampling

During the early 20th century, Marcel Duchamp was creating Readymade Art, which incorporated found objects from daily life that were 'already made'. Under this style artists could choose an object and assign to it, at will, a new identity, utility, or context. The notion of using found sounds had been formalized by a composer named John Cage, a collaborator of Duchamp's, who framed sound as a medium that could also be removed from its original context and assigned a new one. Cage began questioning musical structures and experimenting with electronic sound and new musical media to fill the empty rhythmic and tonal spaces of his compositions (Cage, 1961). In defining noise as non-musical sound and introducing noise into his compositions, Cage believed that "the use of noise to make music will continue and increase until we reach a music produced through the aid of electrical instruments which will make available for musical purposes any and all sounds that can be heard" (Chadabe, 1997, p. 26). Sampling devices would eventually verify this prediction. With the engineering skills and equipment of electronic music pioneers, Bebe and Louis Barron, John Cage composed Williams Mix in 1952 as part of Paul Williams' collaborative music project, titled Project for Music for Magnetic Tape. For this piece Cage recorded over 600 found sounds that he categorized as city sounds, country sounds, electronic sounds, manually produced sounds, wind sounds, and small sounds (Ross, 2007, p. 402). Williams Mix and other pieces that were a part of the project helped lay the foundation for the innumerable possibilities of composition with recorded sound. These figures redefined the role of the composer in music and the use of electronic means of composition, foreshadowing the widespread practice of sampling in popular music genres like Hip-Hop and Electronic Music.

1.2.1 Musique Concrète

Pierre Schaeffer, a French radio technician and composer, embarked on a study of noise with a group of researchers at Club d'Essai de la Radiodiffusion Television Française, a radio research center he founded in 1942. Interested in the capture and manipulation of sounds, the group produced study-like compositions in 1948, including Études aux Chemins de Fer, which sampled sounds from various acoustic environments. Initially a research study into noise, Schaeffer and his team later branded their practice as Musique Concrète (Concrete Music) in the 1950s with the publication of the article, Introduction a la Musique Concrète. In this document, Schaeffer outlined the different sound sources, tools, techniques and methods used in their musical productions, along with a stress of the aesthetic value of their work, rather than solely research (Palombini, 1993). Schaeffer later formed a musical research group, Groupe de Recherches Musicales (GRM), with artists and composers like Luc Ferrari, François-Bernard Mache, and Michel

Philippot (Chadabe, 1997). Iannis Xenakis emerged from the GRM with a unique interpretation of Musique Concrète and the variety of sound sources he sampled and electro-acoustically transformed. Xenakis framed noise as "complex sound-masses that transformed in time as the result of shifting distributions and densities of small, component sounds" (Chadabe, 1997, p. 34) and sought to re-purpose the spectral properties of these masses to form the structures and ethos of his new compositions. Musique Concrète and the GRM challenged traditional instrumentation, conventional tonal theory, and prescribed relationships between composition and performance by "producing, recording, and transforming sounds, disconnecting them from the perception of their original sources" (Palombini, 1993, p. 18). During this era of music technology, a pattern across music began to emerge, where a core practice of isolating sounds from their contexts and fine-tuning their parameters over-arched across many artists' practices.

1.2.2 Electronic Music in Germany and Japan

While France saw an emergence of music technology, studios in Germany were also conducting studies into sound, led primarily at the Westdeutscher Rundfunk (WDR) studio in Cologne. Werner Meyer-Eppler, a lecturer at the Institute of Phonetics at Bonn University, led conversations around music technology and Serialism, first established by Arnold Schoenberg as a 12-tone musical system in response to trends in chromaticism. As twelve-tone systems sought unprecedented control over sound, technology emerged that allowed composers to experiment with a growing set of musical parameters like its playback speed and direction, attack, decay, timbre, and pitch. During this movement in Germany, Karlheinz Stockhausen used sine wave generators, filters, oscillographs, and tape machines to study and organize sound. In 1956, Stockhausen composed Gesang der Jünglinge, which sampled the voice of a boy reading a passage from a book. In other pieces like Telemusik and Hymnem, Stockhausen continued to use tape-recorded sounds that he electronically processed, edited, and mixed together. The WDR studio remained active in producing a number of studies into recorded sounds and composers who looked to experiment with this electronic medium.

Another hub of music technology during the 1950s took place in Japan, when the Jikken Kobo experimental workshop and Japanese Broadcasting Corporation (NHK) opened studios to allow artists experimental domain over sound. Here, again, the idea of recording sounds and classifying them by their properties occurs. In 1953 X, Y, Z by Toshiro Mayuzumi made use of sound recordings from distinct categories of sources including mechanical, natural, and musical. Later in the decade, artists like Yori-Aki Matsudaira and Toshi Ichiyanagi drew upon this collage style of composition. All of this experimentation in music and mixing of styles and techniques coincided with the political opening of nations around the world, promoting an unprecedented mixing and assimilation of global cultures and their sounds (Loubet, 1997). The time period during the 1950s and 1960s saw studios around the world collaborating and opening their doors to electronic mediums in art, which helped liberate sound from conventional practice and theory. This era of experimentation and integration of technology foreshadowed the sampling culture that was adopted in underground music scenes of the late 1970s when Hip-Hop and Electronic musicians also began incorporating found sounds in their compositions.

1.2.3 Turntablism and Sampling in Hip-Hop

By the 1970s, the technocratic ideals of long-Industrialized nations had well-infiltrated mass consumer audiences, with electronic technology touching many facets of society and its production of culture. Music was no exception, shown by electronic music movements emerging at various hubs around the world in the decades leading into the 1970s. Early forms of sampling inspired research and development by consumer electronics companies who began supplying markets with digital sampling devices by the 1980s (Harkins, 2019). This opened music production and composition to larger markets. The Hip-Hop market was one of them. Sampling culture in Hip-Hop, however, popularized the style of embedding prerecorded musical works into new recordings, as opposed to mainly environmental sounds.

Hip-Hop music was born on August 11, 1973 in the recreation room of 1520 Sedgwick Avenue, where DJ Kool Herc and his sister, Cindy Campbell, threw a party that played music in a brand-new style (Wheeler and Bascuñán, 2016). At this party, DJ Kool Herc presented a technique of using two copies of a record and turntables to repeatedly play specific rhythmic sections of the records, which they referred to as breakbeats. A four-bar section when all of the instruments drop out, save for the drums, in Amen Brother by The Winstons was one break beat that frequented early Hip-Hop parties. According to Whosampled.com, this breakbeat was sampled (embedded) in over 4,000 songs, ranging in genres from Hip-Hop to Electronic music. Funky Drummer by James Brown, Apache by the Incredible Bongo Band, and Good Times by Chic hold a similar regard in Hip-Hop history, with their breakbeats being sampled in thousands of songs since their releases in the early 1970s. The Chic sample was used in Rapper's Delight by the Sugarhill

Gang, which was the first Hip-Hop song to be recorded and commercially sold. The success of this record cemented the Sugarhill Gang as international celebrities and is widely regarded as being the first Hip-Hop song to reach success in mainstream audiences.

Grandmaster Flash is a figure in Hip-Hop who refined the breakbeat technique by paying more attention to the tempos of the songs that he was looping to create seamless transitions between breakbeats. Grandwizard Theodore, another pioneer in this art of turntablism, introduced scratching techniques to Flash's mixing strategies. Other DJs around New York adapted this style and injected their own mixing styles. By the early 1980s, Grandwizard Theodore and the Fantastic Five, the Cold Crush Brothers, and Grand Master Flash and the Furious 5, had risen to fame as the most renowned Hip-Hop groups in New York. While all of these groups promoted their individuality, the practice of turntablism and the art of incorporating previously recorded songs into their new creations over-arched across their music and Hip-Hop communities.

DJ Marley Marl is known for ushering in a new era of Hip-Hop production by venturing beyond the pre-programmed drum sounds on devices like the Oberheim DMX and Roland 808. Marl was known for programming instrumental samples from popular breakbeats onto his sampling machines. Through internships at radio stations and studios, Marl was able to experiment with the Fairlight CMI emulator and E-Mu sampler devices, eventually purchasing a Korg SDD-2000 delay with sampling capabilities as one of his first samplers Marl (2014). With the short sampling times of these early devices, drums were a natural instrument to sample due to their quick attack and decay. Marl made use of sampling machines to create many of his first beats from drum samples taken from James Brown records and Impeach the President by the Honey Drippers. However straightforward Marl's idea was to sample individual drum sounds to create new rhythms, it had not been done before and showed listeners the limitless creations that were possible with sampling. Previous to his technique, many producers reused the same breakbeat loops that populated the Hip-Hop community or used the easily recognizable sounds on drum machines. DJ Marley Marl's productions influenced many other producers after him who expanded on sampling in Hip-Hop.

During the late 1980s and early 1990s, characterized as Hip-Hop's golden era, many new artists, producers, and DJs emerged. Since the legal world and cultural policy had not fully caught onto the commercial success that Hip-Hop would become, producers enjoyed a period of unrestricted sampling and creative freedom (McLeod and DiCola, 2011). Groups from this era include Wu-Tang Clan, De La Soul, A Tribe Called Quest, Public Enemy, Eric B. and Rakim, and Boogie

Down Productions, all of which included extensive amounts of sampling throughout their discographies. Known for his diverse array of sample sources, J Dilla emerged slightly after the height of the golden era, producing artists like Talib Kweli, The Pharcyde, Common, and Mos Def. According to WhoSampled.com, J Dilla sampled over 2,000 songs in his discography. Other sampling producers like DJ Premier, the Rza, and DJ Shadow also evolved within sampling culture during this golden era. In the 2000s, despite the storm of litigation that swept over Hip-Hop's sampling practice, the art remained an integral part of music production, with producers like Just Blaze, Kanye West and 9th Wonder maintaining the sampling style in the music of mainstream audiences.

1.3 Early Sampling Technology

All within the last quarter-century of the 1800s, devices like Alexander Graham Bell's Telephone, Thomas Edison's Phonograph, and Valdemar Poulsen's Telegraphone made it possible to transmit sound across wire, store sound on physical mediums, and improve the fidelity of sound recording and reproduction (Daniel et al., 1999). With advancements across acoustics, microcomputing, and signal processing, artists and engineers took to shaping movements like Musique Concrète and Electronic music, which gave way to Hip-Hop and EDM later on. At the root of these arts, were the technologies that supported unprecedented insight into sound and control over its form.

Predating digital sampling and recording, Mellotron produced the MKI in 1963, which was the first commercially available tape sampler. This device took the appearance of a piano and reproduced pre-recorded sounds on magnetic tape, upon striking a key. Drawing parallels to the Mellotron series in fidelity by early critics, the Fairlight Computer Musical Instrument (CMI) was regarded as the first digital sampler to allow users to record and store sound libraries on 8-inch floppy discs (Harkins, 2019). Although the CMI required a user's command-line interaction with computers, it helped shape popular music of the 1980s and was used in one of the first Hip-Hop tracks, Planet Rock by Afrika Bambatta and the Soul Sonic Force, which sampled Kraftwerk's Trans-Europe Express.

In 1985, E-Mu Systems released the SP12 sampling percussion machine, illustrated in figure 1. This 12-bit drum sampling machine included four memory banks that could store eight sounds each. The drum machine came pre-loaded with stock sounds and could be loaded with any sample a producer desired. With its versatility and price, the SP series reduced production costs for many musicians and expanded their musical repertoires. This machine revolutionized sampling

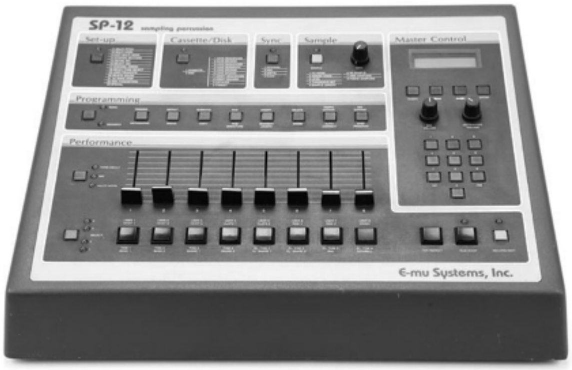

Figure 1: E-Mu SP12 from Harkins (2019)

culture, bringing the technology to an audience who did not require computer programming knowledge or massive production budgets to make music. Despite the SP12's and, it's successor, the SP1200's modernity and versatility at the time, they still placed creative restrictions on the user and are considered to have produced low-fidelity and raw sonic experiences, compared to the devices of today. The E-Mu SP1200, along with the Ensoniq EPS 16+ and ARS-10, were all sampling devices that were used on Wu-Tang Clan's monumental winter-1993 work, Enter the Wu-Tang (36 Chambers) (Sfirse, 2019). Besides the lyrical and thematic content of the work, 36 Chambers is also highly regarded for its production style which was crafted by its producer, The Rza, who had a taste for Kung-Fu movie samples and unique processes to obtain those samples from Videocassette Recorder (VCR) machines. Since the VCR machines' audio reproduction components were incompatible with sampling devices, The Rza introduced a series of adaptors to route the desired audio signal from the VCR through a separate mixer and finally, into the sampler. This makeshift sampling process combined with the analog recorders and magnetic tape of VCRs and The Rza's lax regard for audiophile-quality engineering produced a raw sound in 36 Chambers, which has since appealed to millions of fans around the world.

After the success of the Fairlight and E-Mu Systems devices, a Japanese consumer electronics manufacturer partnered with Roger Linn to release the Akai MPC 60 in 1988. Previously, in 1980, Linn experienced success with the Linn Electronics LM-1 Drum Computer, which differentiated itself from contemporary drum machines by featuring samples of real, rather than synthesized, drums. Using Linn's design and engineering, the Akai MPC 60 also reached commercial success because of its differentiation and versatility. It featured a longer sampling time than the SP-12 and a higher sampling rate of 40kHz (Boardway and Laughton, 2017). With other features like live programming, swing, and MIDI capabilities, it became immensely popular for a range of producers. DJ Shadow, for example, used the MPC 60 as his sole instrument on his 1996 album Endtroducing, which was produced entirely of samples. Until today, the Akai MPC series remains in production, with many other manufacturers also sharing the digital sampler market.

In 1814, German author, E.T.A. Hoffmann wrote a fictional story titled Die Automate (The Automata, in English) in which the protagonist describes an instrument that could "observe closely, study minutely, and discover carefully that class of sounds which belong, most purely and strictly, to Nature herself" (Hoffman ref). The protagonist then describes this mechanical system being enclosed in an instrument that could be played. Playing technology that stored natural sounds as an instrument had come to fruition with early music technologists and composers dating back to Musique Concrète and early tape music. By the 1980s turntablism and digital samplers expanded these studies and experiments with sound to a wider audience of musicians who were given access to a range of technologies that required less engineering and signal processing skills than the equipment used by the GRM and similar 1950s music technology laboratories. The interaction between music and technology has been apparent and is integral to the conversation about sampling. Similar to the computer science and digital signal processing (DSP) advancements that enabled sampling culture, automatic sample detection also relies on such studies into sound.

1.4 Content-Based Audio Analysis

Music Information Retrieval (MIR) is a branch of music technology that is concerned with representing sound and processing it as a signal in novel ways that can be understood and learned by computers. MIR has given way to a range of applications that involve audio analysis. For example, cities today possess machine learning infrastructures that provide real-time information to key stakeholders about sources of noise pollution in their municipalities. This information is used to plan noise mitigation initiatives. As another example, many commercial products and applications rely on voice activation and speech recognition to perform actions on verbal command. At the root of these MIR tasks involves combinations of digital signal theory, computer science, and psycho-acoustics to model human perception of sound and extract information from audio signals. The potential for speed and accuracy over humans in analyzing and recognizing samples in query songs serves the goal of an automatic and irrefutable sample detection system.

1.4.1 Time/Frequency Representations of Audio

Representing sound in the frequency domain is an essential step for computers to begin to highlight patterns in a signal's behavior. In digital signal theory this representation is made possible when a continuous signal (for example, a soundwave) is discretized through a process which is also known as sampling. This kind of sampling, as opposed to sampling in the Hip-Hop context, represents a time-analog signal as a digital one by using the Nyquist theorem. The representation of discrete signals in the time domain, illustrated by Figure 2, can be used to construct its frequency domain representation.

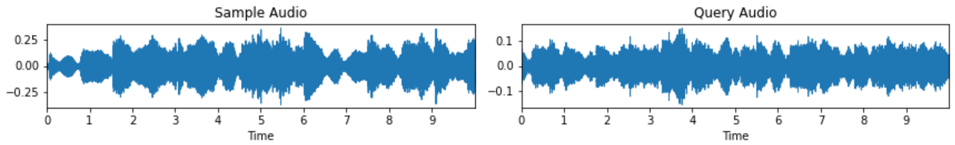

Figure 2: Time-domain representation of two songs, where a sample relationship occurs

From the spectrograms of two songs with a sample relationship, illustrated in Figure 3, patterns within the frequency content of the two signals over time start to emerge. In this example a Short-Time Fourier transform (STFT) was performed on those two pieces of audio, to produce their representation in the time-frequency

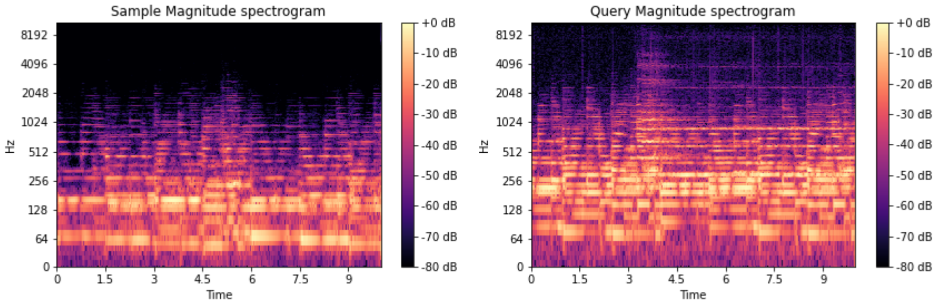

Figure 3: Magnitude spectrograms of two songs, where a sample relationship occurs

domain. The STFT uses the Discrete Fourier Transform (DFT), introduced by Joseph Fourier in 1822. The DFT, described in Equation 1, transforms a discrete periodic signal from the time domain into the frequency domain. Known as a Fourier series, the signal f(t), can be represented as a constant, a_0, and the sum of sinusoids at different frequencies, k and their coefficients a_k and b_k.

$$f(t) = \frac{1}{2}a_0 + \sum_{\mathbf{k=1}}^{\infty}(a_k cos 2\pi kt + b_k sin 2\pi kt) \tag{1}$$

The coefficients of the Fourier series can be achieved by taking the sum of the product of a signal, x(n), and an analyzing function between two finite points in time, shown in Equation 2. Using Euler's formula, the analyzing function, $e^{-j2\pi nk/N}$, can be expressed in complex notation as a sum of its real and imaginary parts, as expressed by Equation 3. With this complex number, magnitude and phase information can be extracted using Pythagorean theorem and an arc tangent, respectively. Ultimately, the Fourier Transform results in the coefficient X(k) for each discrete frequency, k.

$$X(F) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N} \tag{2}$$

$$e^{-j2\pi nk/N} = cos(2\pi nk/N) + jsin(2\pi nk/N) \tag{3}$$

As a final point in the conversation of time-frequency analysis that allows for the spectrograms in Figure 3, the STFT adds a slight modification to the DFT. During the STFT, in order to understand the frequency content of the signal across time, the signal is broken into overlapping windows of time, where the DFT is performed on each window Müller (2015). To accomplish this, the original signal is first multiplied by a window function, w(n), as described in Equation 4.

$$X(m,k) = \sum_{n=0}^{N-1} x(n+mH) * w(n) * e^{-j2\pi nk/N} \tag{4}$$

The type of window function, the window length, N, and the overlap (hop) size, H, are the main parameters of the STFT that affect the time and frequency resolutions of the STFT spectrogram. The result is DFT coefficients, X(m,k) at each point in time, m, and discrete frequency, k. To determine samples in query songs the STFT algorithm was implemented as a first step to ultimately determining similarities in their frequency content, since the query song contains content from the sample.

1.4.2 Harmonic-Percussive Source Separation

Source separation is a topic in MIR that looks to identify structures and layered components of music, which is framed as complex mixtures of sound. The goal of source separation is to decompose this complex mixture of sound into its individual components by leveraging its time-frequency information, represented by a spectrogram, for example. In blind audio source separation, the number of sources is unknown. To combat that challenge, the audio decomposition process can leverage previously known musical information or assumptions about how certain components might behave or appear in the complex signal.

Harmonic-percussive source separation (hpss) was one of the first audio source separation tasks that looked to identify harmonic and percussive sources from a complex audio mixture. This study uses knowledge that percussive sounds appear as transient vertical elements in a spectrogram, covering a broader frequency range, while harmonic sounds appear as horizontal elements in a spectrogram, clustering around certain frequencies. Fitzgerald (2010) described a technique to emphasize the horizontal elements of a complex mixture's spectrogram to arrive at its harmonic spectrogram, while emphasizing the vertical elements to arrive at its percussive spectrogram. This emphasis on either component is achieved with the use of a cost function that helps estimate the harmonic spectrogram and a percussive spectrogram. The cost function is optimized so that when the two resulting spectrograms are summed, they approximate the original spectrogram of the complex mixture, containing both its harmonic and percussive components. The two harmonic and percussive spectrograms could be used as masks to isolate either component from the original signal. Fitzgerald (2010) also proposed a system based on median filtering. This approach viewed harmonic components as outliers in a percussive spectrogram and percussive components as outliers in the harmonic spectrogram, which are stable frequency components across time. Median filtering is a smoothing technique that can remove outliers from a signal by assigning samples of that signal median values from its region, of a given size. Filtering a signal over time or frequency can help remove outliers over time or frequency. Figure 4 illustrates an implementation of a median-filtering based hpss algorithm on an audio mixture that included both types of sources. Since query songs in automatic sample detection can include harmonic and/or percussive components, hpss has the potential to identify samples of harmonic or percussive nature. Including this process in the discussion of automatic sample detection is motivated by the examples of breakbeat style Hip-Hop songs, described in the introduction. In this early style of sampling, drum-based audio was primarily sampled from songs in the query tracks. With this prior musical information, hpss can be used to first isolate the percussive elements of the query track, before conducting similarity analysis between its percussive components and the percussive sample. Alternatively, if the sample was known to be more harmonic, hpss can similarly used in sample detection to separate the harmonic components of the query songs before further processing and similarity analysis.

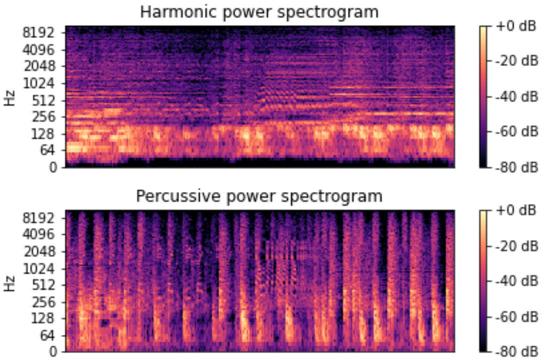

Figure 4: Hpss of a complex audio mixture

1.4.3 Non-Negative Matrix Factorization (NMF)

Some areas of neuroscience and music cognition study human perception of audio signals as the perception of their individual parts (Lee and Seung, 1999). This suggests that humans can separate complex mixtures of sound that they receive into their individual sources. In machine learning and linear algebra, Non-Negative Matrix Factorization (NMF) looks to model this parts-based representation of a signal that psychologists and neuroscientists suspect the brain to do naturally. NMF is a matrix decomposition technique that reduces a non-negative matrix into non-negative matrices of smaller dimensions that, when multiplied, approximate the original matrix. The non-negative constraint on the underlying components of the signal, using NMF, is based on the interpretation of a negative component or weight in the context of an audio signal's magnitude spectrogram. Here, a negative element could not occur and would be difficult to interpret, in terms of what that component would mean as physical sound. As described in Equation 5, NMF can approximate a (n x m) matrix V as the product of its non-negative (n x r) dictionary matrix (W) and non-negative (r x m) activation matrix (H), where their rank (r) is less than m and n and represent a set number of components in V. Throughout literature, the dictionary matrix, W, has also been referred to as basis functions or templates. In an ideal audio application, NMF results in two matrices that represent the spectral components of the original signal, in its templates, W, while the activation function, H, represents the gain associated with each of those components, over time (Müller, 2015, p. 415). Figure 5 illustrates the templates and activations that were decomposed from a spectrogram, with a rank of 5. Given the non-negative constraint, which fits the characteristics of a magnitude spectrogram, NMF has been used in ML and MIR tasks to decompose matrices of complex audio mixtures into their component sources. In automatic sample detection, a query track can be treated as a complex mixture, containing several sources, including a sample.

$$V \approx WH \tag{5}$$

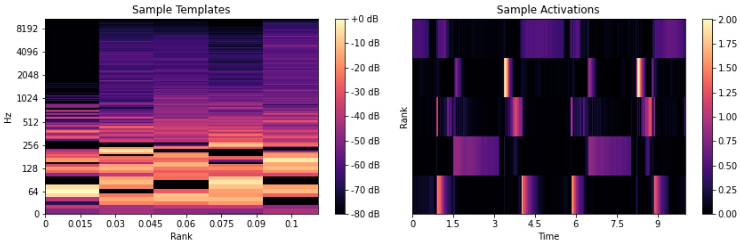

Figure 5: NMF decomposition on a magnitude spectrogram

NMF is an iterative process that approximates the non-negative matrix V by randomly initializing the matrices W and H with non-negative entries and using an update rule to determine how the algorithm should update W and H after each iteration, in order to optimize the objective function. The objective function represents the reconstruction error between V and V_{approx} = WH. A certain number of iterations can be set to determine when the NMF process stops, or a certain cost threshold can be set to terminate the NMF after a low cost has been achieved. Since the optimization of the objective function is an expensive task, standard NMF implementations optimize the function with respect to one matrix, before optimizing the cost function with respect to the other. Lee and Seung (2003) described a basic NMF, including a multiplicative update rule and the Euclidean Distance objective function, d, described in Equation 6. Here, the distance between two matrices, which can be thought of as the original spectrogram and its

NMF approximation, is calculated as the sum of the squared distance between each of their elements.

$$d\left(A,B\right)=\sqrt{\sum_{ij}\left(A_{\mathrm{ij}}-B_{\mathrm{ij}}\right)^{2}} \tag{6}$$

Other objective functions were later proposed by Févotte and Idier (2011), who described a beta-divergence family of cost functions, including the Euclidean Distance, Kullback-Leibler (KL) divergence and Itakura-Saito (IS) divergence, depending on the problem.

1.4.4 Musically-Informed NMF

In order to aid the NMF decomposition in approximating the original matrix, V, a musically informed technique can be leveraged to guide the approximations of the templates and activations, W and H Müller (2015). This technique is also known as knowledge-based NMF constraint in NMF literature.

Partially fixed NMF (PFNMF) is a type of knowledge-based decomposition, where NMF is performed while some or all of the templates, W, or activations, H, are held fixed for some or all of the NMF iterations. The idea here is that the templates or activations that are not held fixed, learn and are updated based on the templates or activations that are held fixed. Wu and Lerch (2015) describe and evaluate this modification to traditional NMF on a drum transcription problem. In this task, the authors used template adaptation to separate the types of drums, including hi-hat, bass drum and snare drum. In order to implement this technique, a database of one shot drum recordings for each drum type was conducted. A one shot refers to a single drum hit. These recording underwent NMF to extract their templates. Next, the pre-computed templates were held fixed during the NMF of the complete drum track to extract activations for the corresponding templates. The goal was that the activations would be adapted to the fixed templates they received. After the activation functions were gathered, onset detection could track when each drum was activated in the recording. The system was evaluated by compiling a groundtruth database of rhythmic sequences, which were annotated with their onset envelopes. The system looked to find commonalities in the groundtruth onsets and the onsets gathered from the PFNMF-derived activation functions. A correct prediction found energy in the activation function where an onset occurred. The system found promising results in their musically informed matrix decomposition. In automatic sample detection, knowledge of the components of at least one of the sources in a query song is known to be the sample song. This knowledge can similarly help inform the decomposition of a query track to better isolate the sample it contains.

1.4.5 Aligning Musical Sequences

Dynamic Time Warping (DTW) is a time-series analysis algorithm that calculates the similarity between two time series by finding an optimal alignment path between those sequences. This algorithm was first applied to sequence alignment in automatic speech recognition to address the impact of speaker variations on speech patterns in time, which tend to be non-linear Sakoe and Chiba (1978). The algorithm can be broken into two steps, which include the calculation of an accumulated cost matrix and the alignment of an optimal warping path.

Given two sequences, X and Y, of length m and n, respectively, the first step in DTW is to create a (m x n) accumulated cost matrix, D. For each element, i, in X, and j, in Y, the cost matrix is computed according to Equation 7. Here, each element in D(i,j) is taken as the difference between each of the elements, i and j, in sequences X and Y, plus the smallest previous value within the three surrounding previous cells in the cost matrix. The accumulated cost matrix, D(i,j), alludes to similarity between the two sequences where the cost (difference) is low.

$$D(i,j) = Dist(i,j) + \min \begin{cases} D(i-1,j) \\ D(i,j-1) \\ D(i-1,j-1) \end{cases} \tag{7}$$

After the cost matrix is calculated, the optimal warping path between the signals is calculated. Constraints on the warping path include boundary, monotonic, and continuity conditions Müller (2015). The boundary condition states that the first and last elements in sequences X and Y should be aligned to each other. The monotonic condition constrains the warping path to move in the same direction in both sequences. In other words, one element in X can only proceed another element in X, if the corresponding elements in Y also behave this way. Lastly, the continuity condition says that no element in X or Y can be skipped. With this optimal warping path, the two sequences can now be aligned and compared, despite distortions in their temporal progression.

Finally, sub-sequence DTW is a modification to the DTW algorithm that looks to align a sequence, X, with a sub-sequence of Y that minimizes distance to X. As in traditional DTW, the first step is to find an accumulated cost matrix, D, between the two sequences. However, the first row and column in D are specially initialized to allow for the algorithm to start at any point in Y, without accumulating any cost. If X exists in a sub-sequence of Y, this step allows the algorithm to skip the part of Y that does not align to X and find the sub-sequence, from all possible sub-sequences, that minimizes its distance to X. From this accumulated cost matrix, an index for the endpoint of X in the sub-sequence of Y can found as the minimum element in the last row of the accumulated cost. The optimal warping path and an index for the start point of X in the sub-sequence of Y can be found by backtracking from the endpoint to the first row in the cost matrix. Finally, another piece of useful information from the accumulated cost matrix is its entire last row, normalized by the length of X. The result, referred to as a matching function and illustrated in Figure 6, illustrates the total cost of the sub-sequence alignment. Local minima in this matching function indicate all end points of X in the sub-sequences of Y.

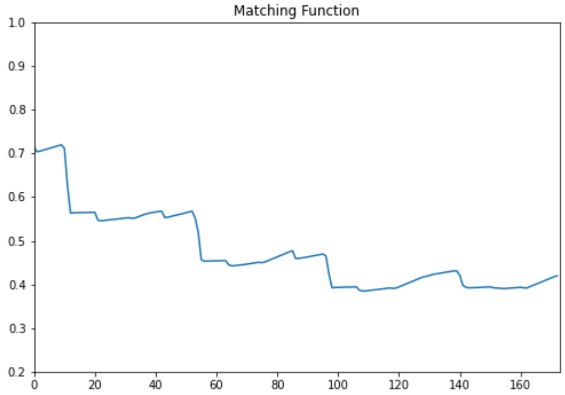

Figure 6: Matching Function of a sub-sequence alignment of a sample that appears 4 times in the query

The use of DTW has been used in automatic sample detection as a similarity measure since it is robust to temporal distortions in the two sequences at hand. This enables the system to recognize samples that have been time-stretched or time-shifted in the query song. It has been found that the accumulated cost and optimal warping path exhibit features that can be further exploited to determine similarity between a pair of query and sample songs.

1.5 Related Work

The range of MIR applications that look to identify similarity in audio are interesting in the discussion of automatic sample detection, which also looks to establish similarity between audio. There are different requirements for each of the applications and tasks. For example, in audio identification, a high level of similarity is required to match two pieces of audio, while a task like cover song detection looks for a more general level of similarity, since the two pieces of audio are not exactly identical. This section reviews some of the relevant MIR studies that pertain to questions of musical similarity.

1.5.1 Audio Identification

One MIR topic that preceded the discussion of automatic sample detection was audio identification, whose ultimate goal was to provide a correct match to a user's audio query. In the early 2000s, Shazam Entertainment, Ltd. developed a fingerprinting algorithm that aimed at robustness against extrinsic variability in the query signal, like external noise, the quality of the capturing device, and the query lengths. In addition, the algorithm aimed at fast computation over a +2 million song database Wang (2003). The method was to extract reproducible hash tokens from query segments and match them with a selection of candidates, which were then evaluated for correctness. Spectrogram peaks, defined as time-frequency points with higher energy than its neighbors in a certain region, were chosen as features to build constellation maps, illustrated in Figure 7, composed of time-frequency coordinates of each peak. The density of the constellation map could be controlled to alter robustness to noise coming through and a desired level of entropy of the dimensionality-reduced representation.

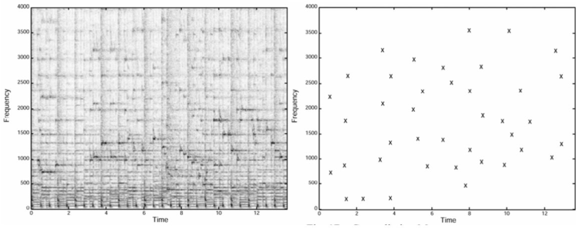

Figure 7: A spectrogram and its constellation map. Adapted from Wang (2003)

Frequency content and distance between peak pairs was used to create hashes, for efficient query lookup in a database of hashes Wang (2003). This spectral peak-based feature accomplished the robustness requirements of a sample recognition system, with high accuracy. In addition, the system reported the ability to identify multiple audio fragments played at once. As reviewed by Cano et al. (2005), the audio identification task has been considered solved by the MIR community. Today, audio identification serves commercial platforms like Shazam and Sound Hound. This task is relevant to this project because automatic sample detection also looks to identify segments of sample audio, despite extrinsic variability in the sample, like distortion or gain. The point of departure, however, is that audio identification is not robust to other forms of variability, like the host of other signal-processing manipulations producers perform on their samples.

1.5.2 Cover Song Detection

An MIR task that is similar to audio identification is cover song detection. In this problem, however, the cover song may differ drastically from the original work, due to variations in performance. Bertin-Mahieux and Ellis (2012) described the requirements of the representation of a query audio (a suspected cover song) to be compact in dimensions, robust to pitch-shifts, and robust to time offsets. Beat-synchronous chroma were extracted from spectrograms and combined with its 2D Fourier Transform to form the feature representation. Chroma features are a feature representation that capture the harmonic evolution of a piece of audio. To compute them, the frequency axis of the spectrogram is converted onto a logarithmic scale based on the logarithmic behavior of a frequencies as they increase by octaves. Next, the harmonics (or partials) of each fundamental frequency are reduced by summing the energy in frequency bands of overtones that are a certain number of octaves from the fundamental. 12 chroma bands are left for each semitone note on a diatonic scale, providing musically-relevant information about the harmonic character of the signal Müller (2015). After being combined with the signal's 2D Fourier Transform and undergoing further dimensionality reduction to a vector of fixed length, the query's feature representation can be compared to the original audio and assessed for similarity.

Cover song detection is useful to this project's research because it provides intuition behind a system that also looks to achieve robustness to differences in pitch, dynamics, and timbre imposed by the performer. However, in practice, the feature representations are too general for the specificity requirements between a query and original audio, in automatic sample detection.

2 Prior Work in Automatic Sample Detection

Automatic sample detection was first discussed and attempted in the MIR community in 2011 by Balen (2011). Since then, the topic has seen other related publications that further address and implement automatic sample detection algorithms, based on the requirements of a state of the art system. This section discusses prior work on the topic. Audio identification and source separation implementations encompass the two main approaches at this problem. The most recent approach to automatic sample detection was described by de Carvalho (2019), which marries these two fields in a two-part hybrid approach.

2.1 Landmark-Based Audio Fingerprinting

The first attempt at automatic sample detection that was discussed and implemented by Balen (2011) and Balen et al. (2012), framed the task as an audio identification problem. In other words, the task looked to identify samples within query audio. That attempt optimized the landmark-based audio fingerprinting algorithm that was discussed by Wang (2003) and implemented by Ellis (2009). The choice of using those resources was to design a system that could withstand various extraneous noise and distortions, which fingerprinting proved to do. The system also needed to be able to identify samples from short appearances in the query. To achieve hashes that satisfied the system requirements, this landmark-based fingerprinting optimization first looked to find the best combination of the following parameters:

- the segment length and hop size of the audio signals
- the number of pairs that are formed for every peak
- the density that the peak picker allows in a block of time
- the peak picker's allowed density on the frequency axis

The mean-average precision (MAP) was taken to gauge the system's accuracy in retrieving the right document(s) (sample(s)) for each query lookup. This measure is calculated as the average precision (AP) for a given number of queries (n), described by Beitzel et al. (2009) and presented below in Equation 8.

$$MAP = \frac{1}{n}\sum_{n} AP_{\mathrm{n}} \tag{8}$$

An experiment was conducted on a groundtruth database of 76 sample relation pairs. Of the 12 queries that returned the correct documents, all of them included samples that had not been pitch shifted or time stretched in the query match. Although the system could not initially handle pitch and time shifted samples, it was able to recognize both tonal and atonal samples. Balen (2011) also experimented with the use of constant q spectrograms to assess how this alternative time-frequency representation would affect the landmark and hashing process. Parameterization of the FFT size and hop size looked to find a balance between time and frequency resolution. Next, the author experimented with a new hash that was designed for robustness to pitch shifts. The results of all of the experiments until this point showed that the regular landmark system, which had returned 12 correct documents, provided the best results. Finally, another method to overcome pitch shifted samples was implemented by repitching the sample's original audio by several factors before undergoing the STFT and landmark hashing process. With 29 correct documents retrieved from the 76 queries, this experiment provided the best results. The retrieved documents included both drum samples and tonal samples and both repitched and non-repitched samples.

This thesis project phrased the sample detection process as an information retrieval task. It was able to evaluate systems over a population of samples that represented a range in the types of samples. Intuition was gathered from the author's error analysis which looked at the types of correct documents that were retrieved and similarities across how they came to be represented in the time and frequency domain. Ultimately the project left questions about landmark-based audio identification as an appropriate method to identify heavily manipulated samples embedded in highly complex query mixtures.

Another audio identification approach was used in an algorithm that is similar to sample detection. Lordelo (2018) developed a system to detect and remove songs in television programs. The similarity to sample detection lies in the fact that the songs would be embedded into a new mixture, containing other sources from the television program's audio, like dialogue and sound effects. The difference to sampling is that the songs would not undergo any pitch shifting, time stretching, or other signal processes in the new audio mixture. Still, a review of this application provided insight and considerations for sample detection, and reiterated the suggestions that an audio signal's features must be robust to degradation and entropic to be recognized by an algorithm. To begin, Lordelo (2018) calculated landmark-based hashes for each song in their database, as done by Wang (2003) and Ellis (2009) in their fingerprinting discussions. To recognize songs in the audio of a television mixture, a landmark-hash was taken of this query and searched over a database of hashes. Since songs in the query audio can have variable gain, as music is faded in or out of scenes, a NMF process was used in a template-matching process to retrieve the signal's gain over time. To remove the audio, scaled versions of the templates were subtracted from the television mixture. The system was evaluated based on its ability to remove audio from complex mixtures. Vincent et al. (2006) proposed a system for performing objective evaluations of blind audio source separation (BASS) tasks, based on the separated components' source-distortion ratio (SDR), source-interference ratio (SIR), sources-noise ratio (SNR), and source-artifacts ratio (SAR). Evaluating their proposed system on an artificially created dataset, the author was able to achieve music removal with robustness to variable gain. While the task of sample detection requires different requirements of robustness, this project's application of a template-informed NMF process provided support to the PFNMF technique. Furthermore, the use of BASS evaluation metrics was an interesting adaptation to suit this problem. For the sake of testing different NMF implementations in automatic sample detection, a similar BASS evaluation could determine their ability to isolate sample sources from query songs.

2.2 Proposing a System to Detect Music Plagiarism

This section describes the prior works that represent the majority of the attempts at automatic sample detection, which are PFNMF-based source separation approaches. These methods frame query songs as a sum of their sources, including a sample source. The evolution of NMF in automatic sample detection is discussed as they provided inspiration and perspective for the set of experiments in this thesis project.

One of the first discussions of sample detection as a source separation task described sampling as a specific type of music plagiarism, where an excerpt from a recorded song is reused in another song. In this paper, (Dittmar et al., 2012) proposed a decomposition process to inspect sampling plagiarism. This approach suggested the use of PFNMF decomposition to approximate the spectrograms, V_s and V_q, of the sample and query songs, respectively, as the product of two smaller matrices. Equation 9 illustrates this approach.

$$V_{\mathrm{s}} \approx W_{\mathrm{s}} H_{\mathrm{s}} \quad \text{and} \quad V_{\mathrm{q}} \approx W_{\mathrm{q}} H_{\mathrm{q}} \tag{9}$$

In these equations, W_s and W_q represent the bases of V_s and V_q respectively, and take the dimensions (n x r). H_s and H_q represent the activation functions of

V_s and V_q respectively and take the dimensions (r x m). As in standard NMF, discussed in the introduction, these activation functions can be regarded as the gain associated with each component in the bases W_s and W_q. The rank value, r, is smaller than n and m. In the proposed method, illustrated by the diagram in Figure 8, the sample song is decomposed using a traditional NMF process, with a rank, r. During the PFNMF decomposition of the query song, however, the basis dictionaries are initialized with a matrix consisting of W_s and a randomly initialized part, W_{q^*}. The dimensions of this basis matrix during the query's PFNMF are (n x (r + k)), where r represents the components of the sample's bases and k represents the components of the randomly initialized part. Throughout each PFNMF iteration, W_s is held fixed and not updated, while W_{q^*} is updated. The activation function H_q, of dimensions ((r + k) x m) is allowed to update and represents gains associated with the bases W_s and additional components in W_{q^*}. If the sample was neither pitch shifted nor time stretched in the query song, correlation between H_s and the part of H_q that is associated with the fixed bases W_s can be conducted in order to determine cases of sampling. Finally, the paper suggested to shift the sample's basis spectra, W_s, along the n axis by a series of factors and repeat the PFNMF process for each of those shifted bases, in order to accommodate pitch-shifted samples. This paper did not offer an evaluation of its proposed system, but provided an interesting workflow for this thesis project.

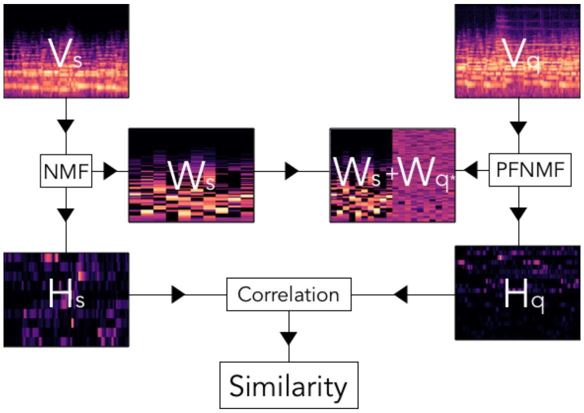

Figure 8: Block Diagram of Proposed NMF System

2.2.1 Implementing PFNMF in Automatic Sample Detection

Following the above proposal to inspect query songs for sampling plagiarism using NMF, Whitney (2013) implemented an NMF-based sample detection system, using High Performance Computing (HPC) to assist in the supervised learning process. This implementation focused on turntable-based sampling, where a change in the sample's speed directly causes a change in pitch.

To begin, Whitney (2013) used auto correlation of the audio signals' onset envelopes to determine their beats per minute (bpm). Using their bpm, each song in the database was resampled to 100bpm. After downmixing each song to a mono channel, in cases where the songs were stereo, spectrograms were calculated and scaled to a log-spaced frequency axis. Also, Whitney (2013) incorporated knowledge about their database's frequency range to adjust the log-magnitude spectrograms to a range of interest between 27.5Hz and 8 kHz. Next, the PFNMF process, as described by Dittmar et al. (2012) was performed for each query-sample lookup pair. The matrix decomposition was configured per Lee and Seung (2003), using a multiplicative update rule and sum of square differences (SSD) objective function at each iteration. The rank value of the sample song's NMF was 32, while the rank of the query song's PFNMF was 64. Finally, after extracting the activation functions of the query and sample songs, a Pearson correlation coefficient was extracted to determine similarity. This coefficient was designed to gauge the strength and direction of two songs' relationship, taking a value between -1 and 1. Whitney (2013) added another step to study information in the dictionary matrices, since they were not discussed by Dittmar et al. (2012). Spectral Flatness was used to determine the noisy and tonal components of a signal, since those components usually exhibit unique patterns in their time-frequency representations. This measure was summarized by Peeters (2004) where a spectral flatness measure (SFM) was taken as the ratio between the geometric and arithmetic means, illustrated in Equations 10 and 11 respectively, where a_k is the energy at frequency band, k, in the time-frequency representation of the signal. Whitney (2013) used the spectral flatness of the components in the basis vectors to distinguish between tonal and noisy components. The sample activations associated with tonal basis vectors were compared to the query activations with tonal bases. The same was done for sample and query activations that belonged to noisy basis components. The goal here was to avoid comparing activation functions that differed in tonality.

$$GeometricMean = \left(\prod_{k \in no.bands} a_k \right)^{\frac{1}{k}} \tag{10}$$

$$ArithmeticMean = \frac{1}{K} \sum_{k \in no.bands} a_k \tag{11}$$

To finally determine a case of sampling, a similarity measure was taken between the query and sample songs, based on the cross correlation between the activation matrices. The evaluation of this system was completed on a database of 10 sample relation pairs. A sample relation pair consists of a query song and a sample song. This project offered promising results to automatic sample detection. However the groundtruth database was limited to only a subset of the types of samples, accounting only for the query-sample relationships where time and pitch shifts are perfectly inversely related. Further, this implementation chopped the query and samples to exact time instances where the sample occurred. Therefore the system's ability to detect samples on an entire song-song basis was also not evaluated.

2.2.2 PFNMF with DTW Similarity Measure

In a more recent implementation of automatic sample detection, Gururani and Lerch (2017) used a groundtruth database of 80 sample relation pairs to train and test their system. This approach used the PFNMF implementation that was described by Dittmar et al. (2012) and implemented by Whitney (2013) to extract the activation matrices for each query-sample lookup. In their pre-processing step, each song was RMS-normalized, reduced to mono channel, and downsampled to 22.05kHz. After magnitude spectrograms were calculated, the sample songs underwent traditional NMF, with a rank of 10, while the query songs underwent PFNMF with a rank of 20. In order to account for time-stretched samples and multiple occurrences of the sample in the query song, the authors of this system used sub-sequence DTW as their measure of similarity, since DTW can withstand distortion in time of two sequences, unlike a euclidean distance or correlation. In order to account for pitch-shifted samples, the fixed basis matrices during the query's PFNMF were shifted by a set of factors that were known to occur in their database. PFNMF and DTW was conducted for each of the pitch-shifted bases from the sample, choosing the pitch shifted basis dictionaries with the lowest DTW cost. Since sample relations often exhibit similar characteristics in their

DTW cost matrices and warping paths, a feature extraction process was conducted for a random forest binary classifier. Gururani and Lerch (2017) included 3 cost-based and 10 warping path-based features for each potential occurrence of a sample in a query, which they counted as a local minimum in the accumulated cost. They referred to the set of local minima as endpoints, since those points indexed locations in the accumulated cost matrix where an optimal warping path of a sample within a subsequence of a query ended. According to Gururani and Lerch (2017), the 13 features included:

1. the minimum DTW cost across all end points
2. the average DTW cost across all end points
3. the standard deviation of the DTW cost across all end points
4. the absolute length of the minimum cost path normalized by the sample length
5. the slope of the minimum cost path
6. the average perpendicular deviation of the minimum cost path from the idealized path, normalized by the length of the path
7. the average slope across all end point paths
8. the standard deviation of the slope across all end point paths
9. the average absolute length of all end point paths normalized by the sample length
10. the standard deviation of the absolute length of all end point paths normalized by the sample length
11. the average perpendicular deviation from the idealized path across all end point paths, normalized by the length of the paths
12. the standard deviation of the perpendicular deviation from the idealized path across all end point paths, normalized by the length of the path
13. the number of end points mapping to this unique start location

To evaluate the binary classifier, the groundtruth data was split into a training set, with 50 sample relations, and a testing set, with 30 sample relations. To save on computing resource, both sets were segmented into 10-song batches, which underwent pairwise query-sample look-ups, rather than look-ups over the entire database. This alteration reduced the number of PFNMF and DTW iterations to 800. Macro and micro accuracy tests were conducted to find song-level occurrences of samples in query songs and the exact location of the samples in the query songs, respectively. For the micro accuracy evaluation, additional annotations were used, which contained time locations for each occurrence of the sample in the query. These included cases of looped samples or multiple occurrences of one-shot samples, and tested the sub-sequence DTW's ability to locate every instance. Gururani and Lerch (2017) used precision, recall, F-scores, and false positive rates to evaluate the macro and micro accuracies.

This project's combination of PFNMF and DTW was successful in detecting samples in query songs, given pitch shifted and time stretched samples. However, with the number of false positive and false negative predictions in the macro and micro tests, there could be improvement to the system's accuracy. Further, the project did not design or evaluate the system for a realistic sample detection problem, where the set of pitch-shifts that the sample underwent in the query were known. In addition, a realistic sample detection task would not have information about the exact time locations of the sample in the query, and so would need to perform the task on a full song-song level. Lastly, this system did not include information from the basis functions, which proved to be successful in the previous sample detection paper at distinguishing the optimal components to compare during the similarity test of the activation matrices. Perhaps this information could be used to reduce the false-positive rate and increase its overall accuracy.

2.2.3 Selecting Query-Sample Candidates

A hybrid approach to automatic sample detection that combined an audio identification technique with source separation was presented by de Carvalho (2019). This project added a pre-processing stage to the PNMF and DTW approach implemented by Gururani and Lerch (2017). In this additional module, a fingerprinting technique was used to first find similarities between query songs and sample songs, selecting the most similar songs as candidates to be passed onto the PFNMF/DTW stage, which conclusively predicted sample relationships. This fingerprinting module used the system developed by Lordelo (2018), discussed previously. The goal in adding this step was to improve the speed of the sample detection algorithm by reducing the number of query-sample look-up pairs within the PFNMF and DTW stage, which proved to be computationally expensive in prior work. Further, by reducing the number of query-sample look-ups, the author hoped to reduce the second module's false positive rate.

To adapt fingerprinting to this specific task, de Carvalho (2019) computed spectrograms, constellation maps, and hashes for blocks of each query song in their database to construct a hash table. Next, the same procedures were conducted on sample songs in order to look them up in the hash table of queries. This look-up accommodated the fact that sample lengths were unknown and could exist over several blocks in the query song. This accommodation was implemented by limiting the distance in time that a subsequent hash in the query song could match to. If the distance exceeded a threshold, the group of selected hashes until that point was terminated, and a new group began. Start and end points of the candidate samples were determined by the time values of the matching hash. If a certain number of hashes did not match between a query and sample candidate, that sample candidate was discarded. Experiments were conducted to determine the optimal parameter values of the candidate selection look-ups. First, the minimum percentage of hashes that needed to be matched between a sample and query song was tested. Second, the block size, in time, from which hashes were taken was tested. Lastly, the maximum distance between query and sample hashes that was tolerable was tested.

The evaluation of this system consisted of two parts. The first part evaluated the fingerprinting module's ability to narrow down the set of sample songs to a group of candidates, which included the one sample belonging to the sample relationship. The second evaluation was used to determine the second module's ability to determine samples and locate them in the query, per Gururani and Lerch (2017). Precision, recall, and F1 scores were used as accuracy measures for both evaluations. According to the literature, the precision for the fingerprinting module was viewed leniently, since the NMF module would ultimately be able to discard any false positive predictions. The recall, while seen as satisfactory, did not accomplish the goal of reducing the computation cost of the second stage. Ultimately this candidate-selection process concluded a lack of entropy in the hashes and thus poor performance on the two-part system by excluding relevant samples as false negative predictions and including irrelevant samples as false positive predictions. While this step was an interesting addition to the current state of the art in sample detection, its design failed to recognize the fact that landmark-based fingerprinting systems are not robust to distortions in time, and so would fail in identifying candidates where time-stretching occurred.

Also venturing from the prior attempt, this project tested the K-Nearest Neighbor (KNN) binary classifier to predict samples based on the extracted DTW cost and warping path features. The KNN classifier works by plotting data points in space and assigning each point a vote, with certain strengths which could vary across the space. A number of neighboring data points, K, need to agree in order for an incoming data point, a query-sample pair, to be classified by the predictor. Using the KNN classifier, the author investigated the number of features that were used in the classification task and the number of local minima in the accumulated DTW cost that were counted as unique sampling instances. The goal was to determine if allowing more local minima hindered the potential length of an optimal warping path, which would then affect the features that were calculated for each minimum. The author found that precision and recall had an inverse relationship in this test. Arriving at worse results than were achieved in Gururani and Lerch (2017), the author found the unbalanced nature of the problem to be of greatest detriment to the precision and recall and speculated about the relevance of all 13 features per sample instance. Still it provided an interesting perspective on automatic sample detection and potentially ruled out certain methods.

3 Methodology and Results

Prior attempts at automatic sample detection showed that musically-informed NMF offered higher accuracy and robustness than the audio identification approach. However, those proposed systems also described areas for improvement and further experimentation. This project evaluated different modifications to the PFNMF process, which was used by Whitney (2013), Gururani and Lerch (2017), and de Carvalho (2019). Further, this project evaluated a baseline and proposed sample detection algorithms on a simulated song-song analysis level, rather than specific regions of sampling, which would be unknown in a real-life sample detection scenario. A discussion of the baseline and proposed experiments follows a look at the ground truth database that was compiled for this research and the set of metrics used to evaluate the experiments. The results of each experiment are included in the experiments' subsections.

3.1 Ground Truth Data

In order to evaluate this MIR task, a ground truth database was needed to establish true and false sample relationships between each of the examples that the proposed system predicted. To build this ground truth database, a request was submitted to Whosampled.com for 1,000 of the top-most user searched sample relations. From the raw crowd-sourced data that was returned, all of the false positive examples that were labeled as sample relations by the platform's users were removed. These mislabeled examples consisted of cover song relationships, lyric interpolations, and melody interpolations. These cases, while representing interpretations of musical influence, did not include relationships where the audio content of the source material was directly embedded into the query tracks, but instead were musically referenced in the lyrics or melodic elements of the new mixture. After this cleanup stage, additional examples were added to the total number of sample relations. While computational resources limited the size of the subset of the data that was eventually used to evaluate the systems, the goal of increasing its size was to contribute additional, cleaned and verified resources for future work.

In its current state, the dataset represents the majority of known sample types, including loops, one-shots, percussive samples, and melodic samples. Furthermore, it includes a diverse range of sample transformations, including pitch shifted, time-stretched, filtered, and otherwise signal processed samples of varying lengths. The fields that were originally included for each of the query and sample songs in the data set were their track name, artist name, release year, record label, destination timing, and source timing. The 'destination timing' field refers to the time stamps (in seconds) in the query song where a sample begins. On the other hand, the 'source timing' field refers to the time stamp in the sample song from where the audio was sourced and used in the query song. Some query songs included multiple destination timings for each of the locations where a sample appeared. In addition, some of the sample songs included multiple source timings, for each of the separate locations where samples were taken from. For these cases of multiple destination and source timings, the cases were split into separate and unique sample relations. For example, the query song 4:44 by Jay Z sampled audio from two distinct locations in the sample song, Late Nights and Heartbreaks by Hannah Williams and the Affirmations. In this example, two separate sample relations were counted. Additional fields were added to include YouTube links for each of the query and sample tracks, along with their filenames, after their audio was retrieved. Table 1 lists select query-sample relation pairs. This table excludes some of the fields for the sake of presentation.

track_name_query	artist_name_query	dest_timing	track_name_source	artist_name_source	source_timing
4:44	Jay-Z	60	Late Nights & Heartbreak	Hannah Williams	204
4:44	Jay-Z	33	Late Nights & Heartbreak	Hannah Williams	1
1 Thing	Amerie	1	Oh, Calcutta!	The Meters	105
330 AM	VI Seconds	0	You & I	One Direction	179
330 AM	VI Seconds	10	You & I	One Direction	218
5% TINT	Travis Scott	0	Cell Therapy	Goodie Mob	2

Table 1: A sample of the ground truth dataset

Lastly, the ground truth dataset was transformed into a matrix to reveal the binary relationship between each of the examples. For this reformatting, each of the columns represented a query song, while each row represented a sample song of a unique sample relationship. The matrix elements were labeled with either 1 or 0, as true or false sample relations, respectively. Table 2 illustrates this format with the same subset of the groundtruth dataset, as illustrated in 1. This transformation was necessary to gather labels for every query-sample lookup.

	4:44_Jay-Z	4:44_Jay-Z	1 Thing_Amerie	330 AM_VI Seconds	330 AM_VI Seconds	5% TINT_Travis Scott
Late Nights & Heartbreak_Hannah Williams	1	0	0	0	0	0
Late Nights & Heartbreak_Hannah Williams	0	1	0	0	0	0
Oh, Calcutta!_The Meters	0	0	1	0	0	0
You & I_One Direction	0	0	0	1	0	0
You & I_One Direction	0	0	0	0	1	0
Cell Therapy_Goodie Mob	0	0	0	0	0	1

Table 2: A sample of the ground truth dataset in binary form

3.2 Evaluation Metrics

The evaluation of the baseline system and experiments in this project looked to determine their ability to detect true query-sample pairs and reject false ones. This evaluation was referred to as macro accuracy by Gururani and Lerch (2017). After predictions were made by the systems, the ground truth data in binary form was used to determine correct and incorrect predictions. Confusion matrices were generated to display true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. TP predictions refer to positive sample relations that were correctly predicted by the system, while TN refers to negative sample relations that were correctly predicted by the system as negative. Perfect scores in TP and TN are the desired outcomes for the system. FP predictions refer to negative sample relations that were incorrectly identified as being positive sample relations, while FN refers to positive sample relation that were incorrectly identified as negative sample relationships. It is the goal to minimize the number of FP and FN predictions. In the event of these incorrect predictions, further analysis and discussion can allude to why the experiments failed in correctly identifying those specific cases. Overall, the relationships of all the predictions were captured in the accuracy, precision, recall, and false positive scores.

3.3 Baseline Experiment

To begin this project, a baseline experiment was conducted using the open-source MATLAB implementation by Gururani and Lerch (2017), which was also used by de Carvalho (2019) and inspired from Dittmar et al. (2012). Implementing a baseline system is necessary in an MIR task to assess and contextualize the results and conclusions reached by the proposed experiments. Further, this practice helped inspire areas of focus in the proposed workflow. This baseline system was chosen since it was the most recent approach to sample detection and since it offered promising results to the task. This system conducted traditional NMF on the sample song, with a rank value of 10, followed by PFNMF decomposition of the query song, with rank 20. From the resulting activation functions, DTW features were extracted and fed into a binary classifier to predict cases of true or false samples.

Since this thesis project focused solely on the system's macro accuracy, some modifications were made to the DTW feature extraction stage of the baseline system, which affect how the results should be interpreted. To review, the DTW stage of the original system sought to extract every unique occurrence of a sample in a query song. To do this, local minima from their activations' accumulated cost matching functions represented the endpoints of every occurrence. For this project's implementation of the baseline's DTW feature extraction process, instead of gathering features for every potential occurrence of the sample in the query, only one set of features were extracted. This set included features for the potential occurrence with the lowest accumulated DTW cost. The original system, conducted by Gururani and Lerch (2017), incorporated micro accuracy in the same test as macro, so at most, only one of the potential occurrences of the sample had to be detected in order to count towards the macro accuracy. Since the baseline here excludes all other potential occurrences, it relied on the one occurrence with the lowest DTW cost to represent the query-sample relation. A larger number of false negative predictions were expected due to this alteration of the experiment, and based on tests conducted by de Carvalho (2019). In their tests, de Carvalho (2019) experimented with excluding local minima in the accumulated cost from the feature extraction and binary classification, and found the system's precision to increase as less occurrences were included, while its recall decreased in the macro score.

Another difference of this project's baseline implementation to consider was that it was tested on a slightly larger dataset than described in the original paper and included a completely different set of sample relations. If the original baseline experiment also included a randomly selected population of samples, then this difference should not impact the results in a major way.

Lastly, this project's baseline system tested query-sample relations over a fixed length of 10 seconds rather than only the regions where the samples occurred. This was done to simulate a real-world scenario, in which those time stamps would not be known in advance. While this is more realistic to the task of full song to song sample detection and should be overcome by subsequence DTW, it inevitably adjusted the problem to test the system's ability to find similarity despite processing irrelevant portions of the query and sample songs. The rest of the baseline experiment, including the NMF/PFNFM and random forest classifier parameters, was conducted exactly as described by Gururani and Lerch (2017).

The subset of data that was taken from the ground truth database included 80 randomly selected training examples and 40 randomly selected testing examples. In an ideal situation, with more computing resources, the system would compare each query song to each sample song in the database. Given 80 songs in the training set, for example, 6,400 comparisons would be made. In practice, this was not done. Instead, the training set was split evenly into eight batches, while the testing set was split into four batches. Each of the batches underwent song-

song comparisons, yielding 800 training predictions and 400 test predictions. This accommodation was also conducted in the original paper to save on resources.

3.3.1 Results and Discussion of the Baseline System

The results for this project's baseline experiment are illustrated in Tables 3 and 4. Of the 40 true sample relations in the ground truth database, only three were predicted to be true. 100% of the true negative sample relations from the ground truth database were predicted correctly. From the results, it appears that the baseline system is highly accurate and effective at rejecting cases of sampling that are indeed not true but highly inaccurate at accepting cases of sampling that are true. These results are extremely far from the original algorithm, but are comparable to results of de Carvalho (2019) who also reported a decrease in recall, as the number of local minima were excluded from the feature extraction.

Accuracy	Precision	Recall	F-Measure	False-Positive Rate
.90	1	.07	.14	.00

Table 3: Evaluation Metrics for the Baseline System

		Ground Truth	
		Sampling	No Sampling
Predictions	Sampling	3	0
	No Sampling	37	360

Table 4: Confusion Matrix for the Baseline System

The perfect precision and high accuracy should be considered against the very low recall, which shows that the system performed very poorly in being able to correctly determine true sample relationships. The system was biased, along the way, to assign negative predictions to the pairs. It is suspected that the feature space was too entropic for the random forest classier to properly make distinctions and split the data in ways that allowed it to make correct predictions. One reason for that could be differences in the types of samples that the system saw, compared to the dataset that was originally used to evaluate the system. Another reason for the lack of order in the feature space is suspected to be the cause of the fixed 10-second windows of analysis between the query and sample pairs. The PFNMF should be robust to highlighting patterns in the activation functions where samples occur in the query. However, since many of the sampling relations in the ground truth dataset included samples that did not last the full 10 seconds, the DTW algorithm may have received true examples with, still, very high accumulated cost. Further, in conducting subsequence alignment, the warping path may have been too constrained, since it is required to warp the beginning and end of the sample song to the subsequence in the query, in creating the optimal warping path. Since 10/13 of the features depended on the warping path, the song-song simulation could have laid outside of the system's robustness.

While the system clearly failed to make correct true positive predictions, an analysis of these cases may shed light on potential merit to the system. Of the three true positive predictions, all included sample relations that were melodic, and two of which were not pitch-shifted. Two of these cases included samples that spanned the full 10 second window, while the one case was a short 2-second sample. This case, however included a portion of the 10-second window that was silent, as the song faded out. While the system performed poorly, a closer analysis of the true positive predictions seem to confirm the explanation that the 10 second analysis window hindered the ultimate similarity score. The false negative predictions included a range of sample types, including harmonic and non pitch-shifted samples of varying lengths.

3.3.2 Results and Discussion of a Second Baseline System

For the sake of discussing the experiments that follow, the baseline system was also evaluated based on the use of a separate DTW similarity score to make sample predictions. This evaluation will be referred to as the second baseline, here. The PFNMF and DTW stages were conducted exactly the same as in the previous baseline. However, instead of extracting 13 features based on the accumulated cost and optimal warping paths before the binary classifier, a prediction was made by the system, solely on the accumulated cost. A threshold was set based on the average minimum cost across a sample of the groundtruth data. If the calculated score fell below the threshold, a true sample relation was predicted and if the score was greater than the threshold, a false sample relation was predicted. The reason for including a single global similarity score was to combat the design element discussed previously, which was suspected to hinder the warping path features from properly representing sample relations. Even though extraneous information from the activation functions would still impact the global similarity score of a true sample relation, areas of low cost (where a sample occurred) would reduce the score against the other false query-sample lookups.

Tables 5 and 6 summarize the results. The higher harmonic mean between precision and recall, represented in the F-measure, show a more balanced distribution of the predictions from before. While the cases of true negative predictions decreased, many more true positive predictions were made. The instances of false positive predictions was significantly higher than previous. The two baseline systems appear to have opposite bias, as this system predicted many more samples, when sample relationships did not occur.

Accuracy	Precision	Recall	F-Measure	False-Positive Rate
.58	.13	.57	.21	.41

Table 5: Evaluation Metrics for the Second Baseline System

		Ground Truth	
		Sampling	No Sampling
Predictions	Sampling	23	151
	No Sampling	17	209

Table 6: Confusion Matrix for the Second Baseline System

While the results were still far from the results described in the original paper, this system offered an ability to detect samples in query songs. Still the high false positive rate was worrying and alluded to a bias, now, towards positive predictions. Analysis was conducted to understand the true positive examples. The same true positive cases that were previously predicted, were also predicted here. It remains the belief that samples that span all or the majority of the 10 second window will naturally be distinguished from the cases that do not, which was further supported by listening to true positive cases. While some cases defied this notion, the majority did not. Furthermore, from a subjective standpoint, the cases that were correctly predicted by the system were easily recognizable by a human listener, as they stood alone or without overpowering sources within the query track, or did not include any pitch shifting or significant time stretching. The true negative examples that were inspected, consisted of a variety of sample types, mainly including vocal samples, drum samples, and heavily buried samples.

3.4 Experiment 1: Fully Fixed NMF

The first experiment that was conducted was based on the second baseline system, which used the PFNMF and DTW similarity score. This experiment was implemented in python and used an implementation where the NMF and PFNMF iterations for the sample and query tracks were conducted with the same rank value. This differs from the second baseline test in that, here, the sample dictionaries were fully fixed to the query dictionaries, disallowing the query to update any additional, randomly initialized, components. This act forced the activations to update fully based on the sample's basis matrix, and no other information. The goal here was to see if placing even greater knowledge-based constraints on the query's decomposition would improve the system's ability to isolate the sample from the query and make a more musically knowledgeable prediction. The DTW scores and predictions were carried out in the same way as in the second baseline. To deal with pitch-shifted samples, however a different technique was used. Instead of pitch shifting the sample's dictionaries by a range of pitch factors and computing PFNMF for each of them, this system shifted the pitches by one factor, based on the difference in tempo. While this action reduced the computation requirements of the system, it excluded a set of samples from the system whose pitch shifts were not a function of their change in tempo.

3.4.1 Results and Discussion of Experiment 1: Fully Fixed NMF

The results for this first experiment, which used the fully fixed NMF method, are described in Table 7, according to the confusion matrix in Table 8. The overall accuracy increased slightly, due to two additional correct TP predictions than before and 58 additional TN predictions.

Accuracy	Precision	Recall	F-Measure	False-Positive Rate
.73	.21	.62	.31	.25

Table 7: Evaluation Metrics for Experiment 1: Fully Fixed NMF

From the results it appears that this system performed very similarly to the second baseline, which would only slightly confirm the theory that heavier knowledge based constraints would be beneficial. Upon closer inspection of the true positive cases, 12 prediction made from this experiment were also made by the second baseline. Given the fact that this system was able to predict the drum and

		Ground Truth	
		Sampling	No Sampling
Predictions	Sampling	25	93
	No Sampling	15	267

Table 8: Confusion Matrix for Experiment 1: Fully Fixed NMF

vocal samples that the second baseline was not, it appears that the two systems slightly favored different characteristics of the query-sample relations. For example, this experiment performed worse on pitch-shifted samples. As explained previously, this experiment used a mechanism to pitch shift the samples that excluded certain examples. This could explain the superior robustness to pitch shifts of the baseline. Both of the systems performed poorly on heavily buried samples.

3.5 Exploring Dictionary Information and NMF Initialization

The experiments described in this section tested two variations of NMF, which approached the musically informed aspect of the query decomposition in slightly different ways. Unlike the baseline experiment and fully fixed experiment, emphasis was placed on the basis matrices of the query and sample songs in determining cases of sampling. This information was not useful in the baseline and fully fixed experiments, since the queries' basis dictionaries would not contain any new or unique information, as they were not updated. To further support this investigation of the basis matrices, Seetharaman and Pardo (2016) described a system that relied on songs' templates and reconstruction error to assess the phenomenon of layered structures and segments in certain composition styles. First, the authors used the basis matrices from traditional NMF to model individual layers, or segments, of a song. They then used these models in a PFNMF approach to approximate other parts of songs, using reconstruction error of the segments across time to determine locations where additional layers were introduced. Since the original basis model could only represent a single layer to the song, reconstruction error would spike when changes to the music, which couldn't fit the model, occurred. This project showed promising results in audio source separation and segmentation and helped inspire the following experiments.

Experiments 2-5 relied solely on similarities in the basis matrices of the query and sample songs to determine cases of sampling. In these experiments, cosine similarity was chosen as a scale-independent measure. Cosine similarity is a measure of distance between two vectors plotted on a unit sphere, taken as the angle between those vectors (Manning et al., 2008). As the angle between the vectors increases, the cosine will decrease, providing a lower similarity score, and vice versa for the angles decreasing between the vectors. Since this angle doesn't include information about the magnitude of the vectors, this measure is scale-independent. In this respect, two basis matrices could still return high similarity, despite relative discrepancies in power. Since this measure is conducted over a vector space, the basis matrices were reshaped accordingly, before the similarity measures were taken.

Similar to the approach in the fully fixed NMF system (Experiment 1) to determine predictions of sampling, here the predictions were also set using a threshold. For each of the NMF variations in Experiments 2-5, cosine similarity thresholds were gathered from the same sample of the groundtruth data used throughout to set thresholds. For every query-sample lookup, if their cosine similarity fell below the threshold, the case would be labeled as a negative prediction and if the cosine similarity stood above the threshold, a positive sample relationship was predicted. This set of experiments held all of the parameters fixed, besides the method of PFNMF that was used.

3.6 Experiment 2: Traditional NMF

This experiment employed standard NMF to both the sample and query songs. As mentioned in Section 3.5, the basis functions where then used to arrive at cosine similarity scores between the songs. Ignoring the merit of a musically-informed approach, this traditional NMF experiment served to confirm the PFNMF practice and offer a discussion in forming the intuition of subsequent experiments. In this light, questions of knowledge-based constraints that are too restrictive were also asked. For example, in the cases of negative sample relations, do these constraints fit the query decomposition too much towards the sample, arriving at a higher similarity? Lastly, this experiment helped start the conversation in this paper about using cosine similarity of the basis matrices as a score to establish macro accuracy, instead of the accumulated cost of DTW measures from activation functions.

3.6.1 Results of Experiment 2: Traditional NMF

The results of the traditional NMF implementations in Experiment 2 are described in Table 9 according to the confusion matrix in Table 10. Although higher in accuracy, this experiment saw a lower harmonic mean between the precision and recall.

Accuracy	Precision	Recall	F-Measure	False-Positive Rate
.56	.11	.50	.18	.43

Table 9: Evaluation Metrics for Experiment 2: Traditional NMF

		Ground Truth	
		Sampling	No Sampling
Predictions	Sampling	20	156
	No Sampling	20	204

Table 10: Confusion Matrix for Experiment 2: Traditional NMF

The slightly higher accuracy over the fully fixed NMF method in Experiment 1 does not necessarily allude to a desired increase in performance. This traditional NMF experiment saw a greater inability to correctly detect cases of sampling. The exclusion of the sample information from the query's NMF is suspected to have contributed to this. Further, the brutality of using an average as the threshold to make predictions also attributed to poor true positive and true negative figures. The extremely high false positive rate was worrying as it exhibited the system's tendency to . The only saving grace in this regard was that the false negative predictions were also high, giving hope that the system was able to distinguish between some cases of true or false sampling relationships and was not completely biased towards either false or negative predictions, despite the unbalanced problem. Of the true positive predictions, four differed from the true positive predictions of the previous experiment. These examples include a percussive sample and three vocal and harmonic samples that the second baseline system was able to detect.

3.7 Experiment 3: NMF with Custom Initialization

This experiment conducted a variation of PFNMF that sits in between Experiment 1 and 2, in terms of the strength of the knowledge-based, or musically informed, constraints. Here, traditional NMF was performed on the sample track, per usual, to extract its set of templates and activations. For the query's NMF approach, however, instead of fixing the sample's templates to the query, the sample's templates were only used to initialize the query decomposition. The intuition was to reintroduce musically informed NMF to the experiment, without imposing such harsh constraints on the query's template and activation function estimates and reconstruction error. This experiment will be referred to as NMF with custom initialization.

3.7.1 Results of Experiment 3: NMF with Custom Initialization

The results from this custom initialized NMF experiment are presented in Table 11 according to the confusion matrix in Table 12. With 62 more TN predictions from Experiment 2's traditional NMF method, this approach was able to slightly improve both precision and recall. This system was able to correct some of the false positive predictions from Experiment 2, without negatively impacting the other predictions. Although far from ideal, this showed an improvement in the system's ability to correctly point out negative sample relationships.

Accuracy	Precision	Recall	F-Measure	False-Positive Rate
.71	.17	.50	.25	.26

Table 11: Evaluation Metrics for Experiment 3: NMF with Custom Initialization

		Ground Truth	
		Sampling	No Sampling
Predictions	Sampling	20	94
	No Sampling	20	266

Table 12: Confusion Matrix for Experiment 3: NMF with Custom Initialization

So far, the first experiment with the fully fixed NMF method provided results with the highest number of true positive predictions. However, experiment 3's lower false-positive rate meant that it was less biased towards positive predictions, yet still managed to achieve better precision than experiment 2's use of traditional NMF. From the true positive predictions, this experiment with custom initialization performed most similarly to the first experiment, which employed a fully fixed NMF. This suggests that the degree to which you constrain the query's NMF delivers only a slight difference in performance.

3.8 Experiment 4: Isolating Query Bases

Ideally specific rank values would be assigned for each song's decomposition based on its complexity and number of sources. However, since fixed values are used, as the number of sources in the songs is unknown, the hope is that in considering the resulting reconstruction error of its components, the decomposition could be more tailored to each unique query-sample relation. This experiment used this intuition to select specific templates to compare after the NMF process.

In the event that this NMF experiment, with its selective process, could better isolate the query's components that correspond most to the sample source it contains, the corresponding templates and activations would improve the system's ability to predict sample relationships. For this reason this experiment included a selection process after the NMF stage, inspired by the use of basis matrices and reconstruction error, described by Whitney (2013) and Seetharaman and Pardo (2016), in their applications. Traditional NMF was conducted on the sample track, while the initialized NMF method from experiment 3 was conducted on the query track. The subsequent selective process chose specific components from the query and sample songs to compare, by taking the squared distance between each query component and each sample component. In doing so, the query components with the lowest distance from the corresponding sample component were selected for the next stage that determined similarity. The goal in doing this was to exclude irrelevant components from the query song, which would be additional sources in the mix, other than the sample. The worry in doing this was that sample information could be distributed over several of the query's components, so by excluding some components from the subsequent similarity measure, valuable information may also be excluded. Another worry was that any sample component had the potential to explain any other query component, including the components not belonging to the correct query or sample. Still, the method was tested and predictions were made as done previously, based on the cosine similarity score and a threshold.

3.8.1 Results of Experiment 4: Isolating Query Bases

The results from this experiment with custom initialized NMF and a selective process are described in Tables 13 and 14. Due to the increase in true positive and true negative predictions, this experiment yielded the best precision and recall so far within the set of experiments that only used template information to make predictions of sample relationships.

This experiment implemented the same custom initialized NMF method as in experiment 3, adding the selective process before finding the cosine similarity of the basis functions. In comparing the specific true positive predictions, they are identical, save for the ten more true positive examples that this experiment

Accuracy	Precision	Recall	F-Measure	False-Positive Rate
.72	.18	.52	.27	.25

Table 13: Evaluation Metrics for Experiment 4: Isolating Query Bases

		Ground Truth	
		Sampling	No Sampling
Predictions	Sampling	21	93
	No Sampling	19	267

Table 14: Confusion Matrix for Experiment 4: Isolating Query Bases

predicted. However, the high false positive rate still alluded to potential lenience towards making true predictions, which may be the result of a higher true positive metric.

3.9 Experiment 5: Differentiating Harmonic and Percussive Sample

This experiment used intuition about sampling culture and knowledge that samples are used for mostly percussive content in the breakbeat query examples. This experiment was inspired by Whitney (2013), who used information about noisy and tonal sections of the basis dictionaries to select specific activations to compare to each other. In their methodology Whitney (2013) conducted spectral flatness measures on the basis matrices to understand the nature of their components. After understanding which components were tonal, only those corresponding activations were taken from the sample and query songs to compare. The same was done for atonal components. This design feature looked to enhance the potential for correct matches. In order to implement this idea, hpss was performed on the query and sample songs to achieve their harmonic and percussive spectrograms. Next, NMF and the initialized NMF process was conducted on each of the sample's and query's harmonic and percussive spectrograms. The dictionaries from the sample and query's harmonic spectrogram were compared using cosine similarity. The same was done for their percussive spectrograms. The higher cosine similarity from the two calculations was used to determine if the sample was either of percussive or harmonic nature, and was taken as the similarity measure. A threshold was used to predict sample relations or not.

3.9.1 Results of Experiment 5: Differentiating Harmonic and Percussive Sample

The results from this hpss process are described in Tables 15 and 16. It did not produce a better accuracy than the previous selective process in experiment 4. In examining the true positive predictions, they were the same predictions made by the previous system.

Accuracy	Precision	Recall	F-Measure	False-Positive Rate
.65	.13	.47	.21	.32

Table 15: Evaluation Metrics for Experiment 5: Differentiating Harmonic and Percussive Sample

		Ground Truth	
		Sampling	No Sampling
Predictions	Sampling	19	118
	No Sampling	21	242

Table 16: Confusion Matrix for Experiment 5: Differentiating Harmonic and Percussive Samples

3.10 Discussion

Exploring different techniques of using the sample song to inform the query song's decomposition were motivated by the goal of achieving a source separation of the query so precise, as to perfectly expose and isolate the sources it contains that belong specifically to the sample song. In doing so, the matter of determining similarity to the sample song would yield more accurate results. The experiment that produced the best results, according to the harmonic mean and false positive rate, was experiment 4. This suggested that the initialized NMF process, along with a selective process provided the most true positive predictions. The results also suggested that there is information in the dictionary matrices worth exploiting to determine cases of sampling, and thus should be considered in the similarity score.

All of the true sample relations were detected by at least one of the experiments. This distribution of true positives could raise concerns that the systems are random in their predictions. However, many of the systems clustered around the same set of true query-sample relationships. These relationships tended to be very easily recognizable by a human listener, although they were still very diverse in their type.

4 Conclusion

Based on conclusions gathered from the results and discussions of each experiment, many more improvements can be made to increase the accuracy and precision of the systems for future iterations of automatic sample detection.

To deal with pitch shifted samples, a bpm counter was used to determine the tempo of both the query and sample tracks, shifting the sample track by a factor that was equal to the difference in tempo. This approach was based on the original method of pitch shifting in tape or vinyl samples, whose pitch was determined by playback speed of the sample song in time. However, some songs in the database included sample relations that defied this assumption. There were query songs that pitch shifted the sample by more or less than the difference in tempo and there were also query songs that pitch shifted the sample without time-stretching it. Modern digital production tools allow for these time-independent changes in pitch. The cases that defied the original assumption ultimately resulted in musically misinformed NMF procedures for the baseline and experiment with initialized sample dictionaries. Further, when it came time to find the cosine or DTW-based similarity of the songs, the dictionaries that resulted from these cases did not exhibit the same amount of energy across their frequency bins, as they were not properly pitch shifted.

Sample detection is a difficult task because of the diversity in their appearance. It may be concluded that a different algorithm would be needed for each type of sample. For that reason, a more focused approach would be desirable in the future, evidenced by the superior results of the prior works that had more narrow focuses. In prior works, systems were built to address specific sample types according to the ones represented by the training and testing databases that were assembled. Given the diversity in sample types and each system's robustness to specific sample types (for example, break-beat style samples or pith shifted samples), an ensemble method could be used in future work. Here, several different NMF methods could be used to inform the ultimate prediction of the system. In addition, a preprocessing step could be introduced to determine the potential type of sample, adjusting the feature-selection process and sample detection approach according to the specific type of sample. Including more annotations about the type, or style, of sample in the ground truth database could aid this implementation. Doing so would also help conduct better true positive, true negative, false positive, and false negative analysis, highlighting the strengths and weaknesses of future work.

Many producers are able to receive stems of their sample's recordings. Stems are the individual audio tracks, corresponding to each source within the sample mixture. Upon further inspection and research on produces' techniques of some examples in the database, it was revealed that stems were used in some of the query songs' production. In these cases, the sample song that was used in the system included sources that were not present in the query song. For these cases, alterations to the ground truth database could be made to include specific stems, and not the full master recording, to address these cases.

This project evaluated a baseline sample detection algorithm and several experiments that attempted a more realistic, song-to-song, analysis. The goal in designing the project as such was to prototype and test different theories to the problem and challenge the state-of-the-art of sample detection's robustness to detecting a variety of sample types across broader stretches of query and sample songs. From the results it was revealed that some of the methods were too crude to handle the realism of this problem. Still, this project can be used to guide future work towards a more refined sample detection algorithm or ensemble of algorithms. Furthermore, this project produced annotated python code and a large +1,000 sample relation database, with audio, that is free and open to the MIR community for future use.

References

Balen, J. V. (2011). Automatic recognition of samples in musical audio. Master's thesis, Universitat Pompeu Fabra, Barcelona.

Balen, J. V., Serrà, J., and Haro, M. (2012). Sample identification in hip hop music. *From Sounds to Music and Emotions. CMMR 2012*, pages 301–312.

Beitzel, S. M., Jensen, E. C., and Frieder, O. (2009). *MAP*, pages 1691–1692. Springer US, Boston, MA.

Bertin-Mahieux, T. and Ellis, D. (2012). Large-scale cover song recognition using the 2d fourier transform magnitude. In *13th International Society for Music Information Retrieval Conference*, pages 241–246.

Boardway, C. and Laughton, J. (2017). A brief history of the akai mpc. Accessed 11 February 2020, https://reverb.com/news/a-brief-history-of-the-akai-mpc.

Burkholder, J. P., Grout, D. J., and Palisca, C. V. (2014). *A History of Western Music*. W. W. Norton and Company.

Cage, J. (1961). *Silence: Lectures and Writings*. Wesleyan University Press.

Cano, P., Batlle, E., Kalker, T., and Haitsma, J. (2005). A review of audio fingerprinting. *Journal of VLSI signal processing systems for signal, image and video technology*, 41:271–284.

Chadabe, J. (1997). *Electric Sound: The Past and Promise of Electronic Music*. Prentice-Hall, Inc, Upper Saddle River, NJ.

Daniel, E. D., Mee, C. D. M., and Clark, M. H. (1999). *Magnetic recording : the first 100 years*. IEEE Press, New York, NY.

de Carvalho, L. L. (2019). Processamento digital de áudio aplicado à detecção de samples musicais. Master's thesis, Universidade Federal do Rio de Janeiro.

Dittmar, C., Hildebrand, K. F., Gaertner, D., Winges, M., Müller, F., and Aichroth, P. (2012). Audio forensics meets music information retrieval - a toolbox for inspection of music plagiarism. In *20th European Signal Processing Conference*.

Ellis, D. (2009). Robust landmark-based audio fingerprinting. web resource, available: http://labrosa.ee.columbia.edu/matlab/fingerprint/.

Fitzgerald, D. (2010). Harmonic/percussive separation using median filtering. *13th International Conference on Digital Audio Effects (DAFx-10).*

Févotte, C. and Idier, J. (2011). Algorithms for nonnegative matrix factorization with the beta divergence. *Neural Computation*, 23(9):2421–2456.

Gururani, S. and Lerch, A. (2017). Automatic sample detection in polyphonic music. *18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017*.

Harkins, P. (2019). *Digital sampling : The Design and Use of Music Technologies*. Routledge.

Lee, D. D. and Seung, H. S. (1999). Letters to nature: Learning the parts of objects by non-negative matrix factorization. *Nature*, 401.

Lee, D. D. and Seung, H. S. (2003). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*.

Lordelo, C. P. V. (2018). Automatic removal of music tracks from tv programs. Master's thesis, Universidade Federal do Rio de Janeiro.

Loubet, E. (1997). The beginnings of electronic music in japan, with a focus on the nhk studio: The 1950s and 1960s. *Computer Music Journal*, 21.4:11–22.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, USA.

Marl, M. (2014). Marley marl on queensbridge rise to fame and hip-hop evolution. Accessed 5 April 2020, http://www.youtube.com.

McLeod, K. and DiCola, P. (2011). *Creative License: The Law and Culture of Digital Sampling*. Duke University Press.

Müller, M. (2015). *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer International Publishing AG.

Palombini, C. (1993). Machine songs v: Pierre schaeffer: From research into noises to experimental music. *Computer Music Journal*, 17.3:14–19.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. *Ircam*.

Ross, A. (2007). *The Rest Is Noise: Listening to the Twentieth Century*. Farrar, Straus and Giroux.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26.1:43–49.

Seetharaman, P. and Pardo, B. (2016). Simultaneous separation and segmentation in layered music. In *ISMIR*.

Sfirse, A. (2019). Engineering the sound: Wu-tang clan's 'enter the wu-tang (36 chambers)'. Accessed 25 May 2020, https://happymag.tv/engineering-the-sound-wu-tang-clans-enter-the-wu-tang-36-chambers/.

Vincent, E., Gribonval, R., and Fevotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469.

Wang, A. (2003). An industrial strength audio search algorithm. *4th International Conference on Music Information Retrieval*.

Wheeler, D. and Bascuñán, R. (2016). Hip-hop evolution. Accessed 9 February 2020, http://www.netflix.com.

Whitney, J. L. (2013). Automatic recognition of samples in hip-hip music through non-negative matrix factorization. Master's thesis, University of Miami.

Wu, C.-W. and Lerch, A. (2015). Drum transcription using partially fixed non-negative matrix factorization. *The 23rd European Signal Processing Conference (EUSIPCO 2015), At Nice, France*.