University of Glasgow | School of Computing Science

Honours Individual Project Dissertation

Automatic Videography of Audio Tracks of Songs

2389622p

April 1, 2022

Abstract

This paper describes an Automatic Videography system, which allows users to generate videos which have imagery corresponding to the audio of a selected YouTube video.

We provide an overview of the past technological pressures on the music industry, accessibility and the field of accessibility studies, and the economics of user-generated content on streaming platforms. Additionally we discuss various similar technologies to that proposed by this paper, and the limitations of state-of-the-art and industry standard speech recognition models when applied to lyrical recognition.

The results of a user survey to evaluate if the system produced videos were able to improve the user experience returned inconclusive. However, users provided valuable insights into aspects of the system that need to be changed or improved within future iterations of the project, and additional research into user attitudes of system behaviour.

Education Use Consent

Consent for educational reuse withheld. Do not distribute.

Contents

1 Introduction

This chapter will outline the motivations that led to the creation of the Automatic Videography tool as well as what we aim to achieve with this system. The Automatic Videography tool enables users to generate videos which have imagery corresponding to the audio of a given song. This is done through identifying keywords, or keyphrases, within its lyrics, retrieving images that represent these key phrases and deciphering the audio, caption information, or video frames to determine when these images should be displayed.

1.1 Motivation

By creating a system that is able to automatically produce visual cues based on the audio content, we can investigate the effects of multimodal information sources in enriching the original song. There is currently evidence to suggest that multimodal mediums can aid in teaching and learning [8]. Additionally, such aids improve accessibility of media objects through translation between sensory categories [20]. A tool such as the one outlined in this dissertation would enable creators to create music videos if they do not have the ability or capital to create a traditional music video manually. This would also allow for user derivative pieces to be created on songs independent of the artist, which is of great benefit in the case that the artist does not have the will to, or is otherwise unable to, create a music video.

This dissertation provides a review of the current limitations of commercially available technologies in musical retrieval, specifically lyric recognition. Such technologies have applications in the indexing of songs, and the automatic generation of subtitling, which links to the important role of technology towards furthering accessibility.

1.2 Objectives

This dissertation revolves around a system which given any YouTube video, which is referenced through its unique hyperlink to that video, can compile and display a video, with imagery that is referring the lyrics of the source video, to the user. In addition to requiring the YouTube hyperlink, the system will also require the name of the song and the artist; this is required for the lyric retrieval as the metadata for a song cannot be reliably sourced, which is one problem with metadata discussed later, in section 2.3.2. The system should be easy to use, through a GUI that is intuitive and provides feedback to the user to facilitate use and error handling. The system must have a way of determining which words should be represented as an image and to determine from when to start and stop the display of said image. The system proposed outlines three different methods to achieve this.

The metrics of success for the system include the improvement of user experience between generated videos and still image videos, as well as showing relevant images in a timely manner. These shall be measured by recruiting participants for a user survey in which each individual will be randomly exposed to a random sequence of either generated or still generic videos of the same songs. They shall provide qualitative feedback on their experience of each video, be it generated or generic still, as to contrast user experience between the two forms of media. Participants will also provide their feedback on how timely and relevant the images that appear are relation to the word spoken. They will also be encouraged to provide feedback on the system through an open question on the system as a whole.

There will also be an exploration on the impacts that such a tool could have if released commercially, by comparing how user-generated content has shaped the musical landscape of YouTube, a service which uniquely capitalises on user-generated content, with other streaming services; this shall include a history of music technologically and socially, and current trends in the interactions between user-generated content and consumers, users acting as creators, and the original artist. We will also explore how streaming and user consumption on YouTube is a great boon economically to the music industry after many consecutive years of loss in revenue, and how it may also ultimately negatively affect the music industry by stifling new creatives and decrease genre diversity. One such issue is difficulties in the remuneration of artists for their works, in part caused by a lack of a standardised database for song metadata to identify the artist that is to be paid. This shall link closely to user-generated content, which is how the product of this system can be classified.

The requirements of this system necessitates performing some form of musical retrieval: the extraction of characteristics of a musical piece, namely the timings of known words of lyrics, known as lyrical alignment. Lyrical alignment of polyphonic music is difficult. Many companies, such as Apple [23], have indicated their interest in an accurate and scalable lyric transcription system, to compute work that is currently done manually by teams of transcribers. We must examine the state of speech recognition, and how the composition of music is not suited to many commercially available speech recognition models; current and on-going research has led to the components that have been used in the implementation of this system.

Another component of this system is the keyword extraction, therefore we shall examine the various keyword extraction tools that are currently available; this shall explore the need for keyword extractors, the limitations of past keyword extractors, and an overview of the extraction tool, Yet Another Keyword Extractor (YAKE), that is used to determine keywords for image candidates.

To fully understand the scope of the work done in this area it is necessary to dissect similar systems that have been created. In this case, there are two different but comparable systems: the first is a photo slideshow generation tool and the second is an image generator based on the lyrics of the song. These two systems will be compared and contrasted with our own system, to point out where our system approved upon them, or what can be taken from these systems to further improve upon our own.

2 Background

The chapter will explore the following: the history and evolution of music, the interactions between music with society and economics, and the impact that user-generated content, which the product of our system would be categorised under, has on the online music landscape. This will also lead into an exploration of works similar to this dissertation project.

2.1 The History and Evolution of Modern Music in Society

Like most things in the modern world, the access to music at this present time is unprecedented through the internet. In previous years, music had to be stored physically in either cassette tapes or vinyl records, and a natural consequence of this is that collections of physical copies had to be bought, borrowed, traded [28], or copied [6].

This changed rather drastically during the mid-1990s as it became increasingly more common for the music industry to distribute music using CDs. With the content digitised, and therefore easily reproducible, it resulted in the increasing prevalence of the Internet music, as it could be be shared online through peer-to-peer (P2P) file sharing platforms, such as LimeWire, Kazaa, and Pirate Bay [6].

These new modes of music distribution were a thorn in the side of the music industry for many years, slashing their revenue. There existed no real solution to this problem until the mid-2010s when music streaming became the most common means through which to consume music. Through these streaming services, users can now uniformly share vast amounts of music and a growing number of people can access this music without owning it [6].

2.1.1 Socialisation and music

Music has the ability to influence the mood of individuals, strengthen social ties, and create shared meanings [28]; allowing the creation of a shared identity between people.

The limitations of its physicality created a culture around the proximity and accessibly of these collections. Due to the comparatively low portability to music today, this necessitated individuals gather in locales with access to shared jukeboxes. In such public spaces music may be democratised "[encouraging] debate, conversations, and negotiations" [28], and the sharing of differing musical tastes brings about camaraderie within and between social groups. The reduced access and greater monetary investment incentivised trading or sharing of physical copies, creating an organic network through which to share recommendations and create social groups. This could also take place in private places of residence, where music apparatus was commonly in communal areas, creating opportunities for spontaneous inter-generation conversion and bonding [28].

Sharing of physical media involved sharing with friends or families, whereas when commenting about music on social media platforms the majority of individuals whom interface with the messages will be 'familiar strangers' [28], with the connection between these individuals being tenuous; this may be through friends-of-friends or through a common interest. Interactions with these familiar strangers through music expands the musical ecosystem, enabling people to socialise, explore, discover, share, and recommend music [28] with more people than they would be able to face-to-face..

The use of then iPods, and the now smartphones, allowed for the reconfiguration of the musical ecosystem due to their versatility [28]. Private technologies such as headphones change the nature of music, as it makes the listening experience more intimate.

2.2 Copyright and Safe Harbour

Piracy has been a major issue for the music industry for many years, starting even before the shift towards digitisation with P2P file sharing sites. Despite fighting to maintain their copyright out of financial necessity, the legal proceedings that were put to those file sharing sites garnered negative publicity in addition to the court cost required. However, these cases ultimately were successful in which Limewire and Napster were required to pay $100 million as compensation, and also forced Napster and Kazaa to become legitimate streaming sites [6].

In order to assist the music industry against piracy, the World Intellectual Property Organisation (WIPO) created the WIPO Copyright Treaty (WCT). The treaty sought to regulate the consumption of digital media and grant record labels the exclusive right to works produced by performers [6]. The treaty would go on to be adopted by EU member states in 2002, and subsequently adopted by the UK in October 2003 [6].

Another piece of legislation which is more controversial within the context of the music industry is that of safe harbour provisions adopted in 2002 by the UK, due to being an EU member state. Safe harbour laws outline the liability of Internet Service Providers (ISPs) pertaining to the types of information that they can transmit, cache, or host which would otherwise infringe on the copyright of an individual(s). These provisions sought to refine the level of financial and legal liability that an internet provider is burdened with under certain conditions. These conditions stipulate that an ISP has limited liability in the instance that they act merely as a conduit to transmissions, and must not facilitate illegal activities but instead be diligent in removal of illegal materials when they have knowledge of such materials [6]. The inability to do so may put a platform at risk of losing its privileges provided under safe harbour. Illegal activity which is relevant in this case would be the sharing of copyrighted materials without the permission of the copyright holder.

2.2.1 User Generated Content and Youtube

Such safe harbour provisions also apply to platforms that host user-generated content(UGC), such as YouTube, granting them unfair advantages over other streaming services, which lies in how it interacts with safe harbour laws. For other streaming services, licences to host music must be negotiated before making that content available. In contrast, YouTube can host content from their users before seeking a licence, at times without the intention to seek it at all [6].

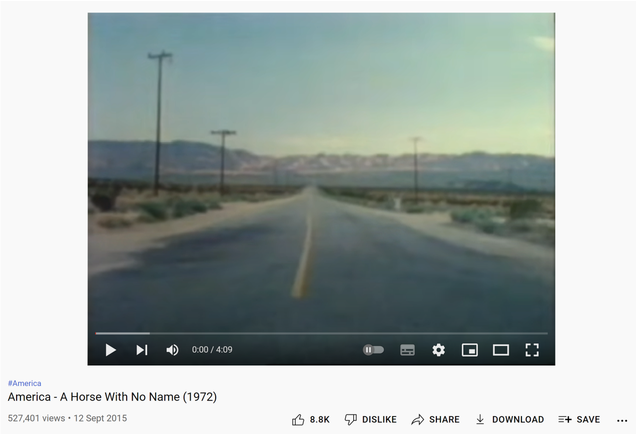
YouTube provides tools which allow copyrights holders can maintain their copyright through two main avenues: Copyright Takedown Requests and YouTube ContentID [17]. The former is a manual webform to be submitted once a video has been identified by, or on the behalf of, the rightful copyright owner. The latter is a probabilistic matching system used to determine if content uploaded has a similar fingerprint to files submitted by copyright holders. This step requires the copyright holder to have first submitted a number of claims previously and for them to have submitted these files manually. Once the claim has been made the copyright holder may take actions on the video, such as blocking the video so it cannot be viewed, run advertisements on the videos to generate money, or monitoring the videos viewership[17].

2.2.2 The Topology of YouTube

The most popular videos on YouTube are those that are in the music genre, with 38.4% of all traffic through YouTube in 2013 was in reference to music and 23-30% of all videos bearing the *Music* tag [29]. This can be attributed to the tendency for users to generate *user-appropriated* content which may range from re-invented to outright copying of the original. In some ways, YouTube may act as an archive for all videos, with music videos in particular it is common that multiple of the same, or similar, copies of the same song are uploaded. However, unlike an archive these copies may be removed by users by deleting their accounts or due to copyright claims [29].

Uploading to YouTube is an uncommon behaviour, as only 11% [29] of all users will upload any videos. All videos that are within the *Music* category can be split into three primary types - *Traditional*, *User-Appropriated*, and *Derivative* [29] - which can be split into a total of 12 subtypes. These subtypes are listed and characterised in table 2.1, and for some examples of video types on YouTube refer to Figures 2.1 and 2.2.

The interaction between users and a music video has been observed to differ depending on its subtype [29]. For example, user-generated content such as *covers* and *parodies* have the highest engagement of the subtypes, supporting the idea that user-generated content sparks engagement. But this is not true for *lyric* and *still* videos which do not receive as much engagement as the authentic *classic music video*, suggesting that perceived authenticity is the driving factor in engaging with the video[29].



#America

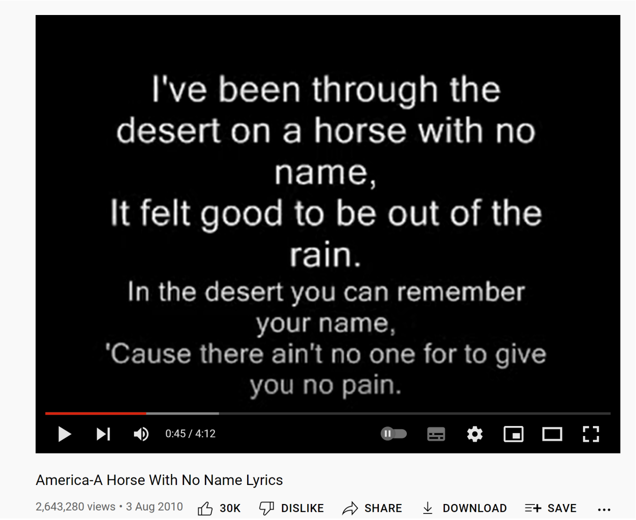
America - A Horse With No Name (1972)

527,401 views • 12 Sept 2015

Figure 2.1: *This figure shows an example of a Classic Music video, which is in the 'Traditional' category of YouTube content.*

User-appropriated content is a subset of user-generated content, and is characterised as content which has been copied but has not been significantly changed to warrant the copy to be considered novel [29]. However, the extent to which a piece of content can be modified yet still be considered a copy is a difficult and contentious topic, and there may never be a clear answer.

Users are more tolerant of some modifications over others. For example, users are more tolerant to changes in audio content than to changes in the video content. However, this is only the case when the change in video content changes what the perceived identity of the song is [29]. In the case that there is a song by an artist that is perceived negatively then this will correlate with an increase in negative engagement, which is not true of the parodies or covers of the same song which do not have the same proportional negative engagement; this is referred to as the '*hater bump*' [29].



I've been through the
desert on a horse with no
name,
It felt good to be out of the
rain.
In the desert you can remember
your name,
'Cause there ain't no one for to give
you no pain.

America-A Horse With No Name Lyrics

2,643,280 views • 3 Aug 2010

Figure 2.2: *This figure shows an example of a Lyrics video, which is in the 'User-Appropriated' category of YouTube content.*

2.2.3 Curation

Due to the homogeneity of content across streaming services, one way in which to differentiate streaming services from each other is in how they curate this content [6], through the creation of playlists. There are three different types of curation: user-curated, editorial and algorithmic, with the main differentiation between streaming services lying in their editorial and algorithmic curation.

Editorial curation is performed manually by the streaming service itself, or by a well-respected third-party. However, due to the human factor, these curation processes can be corrupted through eliciting brides from artists to ensure preferential consideration; allegedly it is still common for playlist curators to promote artists for a price [6]. The algorithmic curation process refers to the use of recommender systems to compile personalised playlists. These however also pose issues for artists in that introducing new music requires 'training the algorithm', [6] in which artists attempt to maximise the traffic to increase their chances of being recognised by the algorithm, which is biased against independent artists. The tendency for editorial and algorithmic curation to draw from a narrow range of music has been noted to cause an 'echo chamber' [6]. This traps people into their own music tastes which are excessively mainstream, starving more niche and newer artists from expanding into the market. This will eventually lead to a decrease in the variety of music from the musical ecosystem, as experimental new artists are unable to make money [6]. Conversely, videos from popular artists will share an audience with other similar popular videos, thereby increasing the viewership of both videos reciprocally; this phenomenon is called the *halo effect* [29].

2.3 Economics of Online Media

Music streaming has been the saving grace for the music industry, allowing the music industry to move from revenue losses to now experience year-on-year growth since 2014 [6]. However, despite streaming services being the main contributor to increasing the revenues of the music industry as a whole, they themselves have not historically been profitable. The combined annual losses of Spotify during the decade prior to 2020 reached €2.62billion [6], with Spotify and other services being kept afloat due to venture capital. The main exception to this trend is in

Primary Types	Subtype Name	Characteristics
Traditional	Classic Music Video	A professionally produced short movie set to the song, and is either uploaded by representatives of the artist or uploaded by a user as a copy of the original.
	Alternative Version	A video with identical audio to that of the Classic music video, however it shows different imagery.
	Live Music	A video copy of a TV broadcast or professional live recordings uploaded without modification.
User-Appropriated	Still	A music video which has only a single image or slideshow of images as its only visual.
	Lyrics	A music video resembling a still video, however the visual contents also include the song lyrics.
	Embedded Lyrics	A music video in which the visual component is that of the classic music video with an overlay of the lyrics.
	Fan Illustrated	A music video in which the visual components have been reinterpreted in a new medium.
Derivative	Cover	A music video in which the audio is that of a different artist than the original song.
	Dance	A video in which the visuals are that of a dance routine, however the audio is usually copied.
	Parody	A video which is meant to be a comical interpretation of the original.
	User-Illustrated	A video is which the original audio is the same, however the visuals are partially, or completely, unrelated.
	Other	A video which falls under no other category.

Table 2.1: *The twelve subtypes of music videos and their characteristics*

2020 during the COVID-19 pandemic in which Spotify received a surge in subscribers totalling 30 million subscribers and doubled its share price, and representatives of Spotify attribute this increase to the loss of live music [6].

2.3.1 Economics of YouTube

As previously stated, copyright claims made against user-appropriated content on YouTube enables the copyright holder to monetise the content. In fact, due to the multitude of exact, or similar, copies of a video that are being monetised, the revenue generated through these copies exceeds that of the original copy [29]. However, in terms of renumeration to the artist YouTube pays less than average per-stream than any other platform [6]. This is common among UCG-sites which often lag behind the market in terms of remuneration, in a phenomenon called the 'value gap' [6]. YouTube retorts this by stating it currently provides 30,000 worth of jobs to the UK economy as well as £1.4 billion towards GDP, and claims to be on course to be the number one source of revenue for the music industry by 2025 [6]. It is, however, unclear if the latter point, of an ad-based system being the basis of the music industry, is conducive to a healthy ecosystem; as streaming on YouTube currently makes up 51% of all streams per year, but only accounts for 7% of all revenue [6].

Another factor which ultimately decreases remuneration for artists across all services, is that

YouTube is free. The homogeneity of the products offered by streaming services creates a perfectly competitive market where consumers can find a perfect substitute from another streaming service if they so choose. This has caused prices for streaming services to become static at around 9.99 independent of currency, without regard for exchange rates [6]. Any increase in price could cause a consumer to change to any other cheaper streaming service, and as YouTube is free this forces all other streaming services to lower their prices. This loss in revenue necessitates that the rate of renumeration remains low across all streaming services. However, YouTube argues that it being free reduces the usage of P2P sites, which are now visited regularly by only 2.9% [6] of consumers, allowing them to monetise what otherwise would have been lost revenue. The attitude of consumers when it comes to renumeration is minimal in that most consumers are unaware of the process and most do not consider ethical sourcing when consuming music [6]; although consumers do not support an increase in renumeration under the current model, they would instead support funding artists directly through various means, for example crowd sourcing websites [6].

Half the revenue generated by YouTube for the music industry comes from ContentID, one of its copyright tools. However, it is important to note that although it has been deemed by stakeholders to be effective and scalable, no proposed system currently is on par with human reporting and review [6]. This imperfection makes it possible to make money through user-appropriated content by making slight modifications to avoid detection which is then paid to the creator of said modified content before a copyright claim has been made against the video [6, 29]. Despite this, people who post user-appropriated content are not looking for monetary gain.

2.3.2 The Problem with Metadata

There are three main components that are transferred in that case the music is licensed to a streaming service, those being the music track, the cover artworks and the metadata (information about the music track itself). Of this metadata, there are also three components which are relevant in music streaming, these being descriptive, ownership, and recommendation metadata. Description metadata refers to information pertaining to the contents of the music track. Ownership metadata refers to information on the creator and record label, and contractual agreement on which the music is licensed. Recommendation metadata refers to tags applied to assist in the function of recommender systems and unlike the other forms this type is not static, with associated tags changing to reflect user behaviour [6]. Ownership and description metadata is created during the songs creation, in which the songs are assigned two unique identifiers. These are called International Standard Musical Works Code (ISWC) and International Standard Recording Code (ISRC), which are used to authenticate and differentiate songs and recordings and to catalogue the proper rightsholders [6].

The ability for streaming services to properly remunerate an artist has been in part due to the inability to identify the artist in question. As songs and recordings are licenced separately, the information is often held by different parties. Therefore the music labels have the ISRC they require to license the song but does not necessarily pass on the ISWC which is necessary to identify the artist, as it may not even be known to them. This is worsened in the case that there are multiple recordings of a single song, all of which should share the same ISWC, which will then have no information available to trace back to the artist [6].

Streaming services also exasperate this problem by restricting the means to provide accurate metadata, whilst also inhibiting the ability to contest incorrect metadata [6]. One such example is the requirement that such corrections must come from the rightsholder or a representative of the rightsholder. This prevents a good actor, which is a third party in this exchange, from assisting the original artist from the impractical task of sifting through and claiming all of their stolen content. Additionally, the consensus among streaming services is to urge legislators to find a solution which ensures that ownership metadata is sufficient to identify the artist [6]. One such solution that has been suggested is to make a public and standard database in which to identify artists.

This problem is relevant to our system because without a reliable way to extract the metadata of the video pertaining to the ownership metadata, which may not even be contained within it, then the best way to determine the song title and artist name is through user input.

2.4 Accessibility

A system capable of increasing the modality of media is connected to the idea of accessibility in that this is one small aspect of making media accessible. The increased prevalence of audiovisual media through streaming has been a driving force in the evolution of the accessibility studies, due to the increased risk of discrimination that it poses. The consensus of what accessibility is and what it should encompass has shifted in the field of accessibility studies, deeming some ideas within as discriminatory whilst they also remain prevalent throughout the field [20]. These ideas stem from two different views: should accessibility be a human right or is accessibility a tool in furtherance of human rights. The former position was reinforced by the United Nations in the passing of the Convention of the Rights of Persons with Disabilities in 2006 [36], then adopted by the UK in 2009, in which they recognise accessibility as a human right, specifically for people with disabilities [20].

To say that accessibility as a human right is to be recognised in some specific group, but not others, could be seen to come into conflict with other human rights. For example, if we consider freedom to be the ability to do something without restriction then this would imply that freedom is fully dependent upon accessibility, which pertains to the ability to access that thing in the first place; this is considered controversial, as to say that a human right is not itself the foundation would require redefining what a human right is [20]. Additionally, human rights are said to flow from the inherent humanity of the individual, therefore to say that some individuals have different human rights may further, or create, discriminatory rhetoric, reinforcing a *ghetto effect* [20].

If we however reframe accessibility as a tool with which human rights can be furthered, but not as a human right itself, then this can complement our current understanding of what a human right is. In accordance to this principle, in 2011 the World Health Organisation defined accessibility as "the degree to which an environment, service, or product allows access by as many people as possible, in particular people with disabilities" [37]; this framing ensures that furthering access for people with disabilities remains the goal, but expands the definition to include additional barriers to entry, such as linguistic and cultural barriers. In either case, breaking these barriers requires translation without loss of meaning.

These different perspectives can be encapsulated under the terms of particularist and universalist accounts [20]. Particularist accounts are similar to the former explanation of accessibility as a human right, where accessibility is reduced to only apply to some groups of people or some types of barriers, and limit media accessibility to translation-based modalities [20]. Universalist accounts, conversely, adopt accessibility as tool to further human rights and does not seek to reduce accessibility to any group or barrier, and neither does it limit media accessibility to only that of translation-based modalities [20]. The basis of the universalist account is that of the *social model of accessibility* in which the interactions the user has with media, or even non-media, objects [20] are the focus. Namely addressing the gap between the user's capabilities and their needs [19, 18].

Within the field of accessibility studies there has been trends over the years towards a universalist-centric view over the particularist-centric views held in the past; this shift is in part due to sections within the particularist accounts being deemed to be discriminatory [20]. It has been observed that particularist accounts become internalised by students despite media accessibility, in these learning critical spaces, as they are taught through the lens of the universalist accounts [20]. This informs a bias against the universalist accounts within the students in which it has been observed to foster restrictive ideas of media accessibility and potential discriminatory perspectives [19].

An additional issue under the particularistic framework, is that it would restrict the understanding of media accessibility to that of translation. Consider the problem of extracting clear audio from a video or soundtrack, i.e. removal of any background noise and enhancement of the intelligibility of dialogue. The importance of having access to clean audio to facilitate comprehension is recognised by scholars, service providers and organisational bodies for disability alike, however this is undermined by the difficulty in creating it. This is not a matter in translation, but instead that of musical retrieval.

When it comes to the creation of derivative media to aid in accessibility there has often been a disconnect between the function of an accessibility tool and the user's needs; this is referred to as the maker-user gap [20]. Other time the user's experience has been integrated into the creation of accessible products, marking a change towards user-centric approaches. These new approaches have opened many different pathways towards research in audiovisual translation, and create a new lens through which to examine the efficacy of conventional audiovisual techniques [20].

However, the recreation of artefacts in media from the ground up is usually deemed an inferior approach in accessible artefact generation, compared to integrating principles of accessibility into the design and creation of the original piece. Additionally, deviations from the original piece, even partially, in terms of aesthetics or author intentions has shown to have a negative affect upon the user experience [20].

2.5 Similar Technologies

In this section we shall examine two previous systems which have aspects which are similar to this project. The first is a system created in 2010 which automatically generates a slideshow based on the lyrics of the song [13] which is fairly similar to captions and lyric extraction methods in the way that the method is line-level-alignment, not word-level-alignment. The three main outlined steps of this system are Candidate Image Retrieval, Image Selection, and Synchronized Playback. The second system was created in 2021, which unlike the first and the system of this dissertation, it produces an image based on the mood of the song lyrics, suitable for an album cover [2].

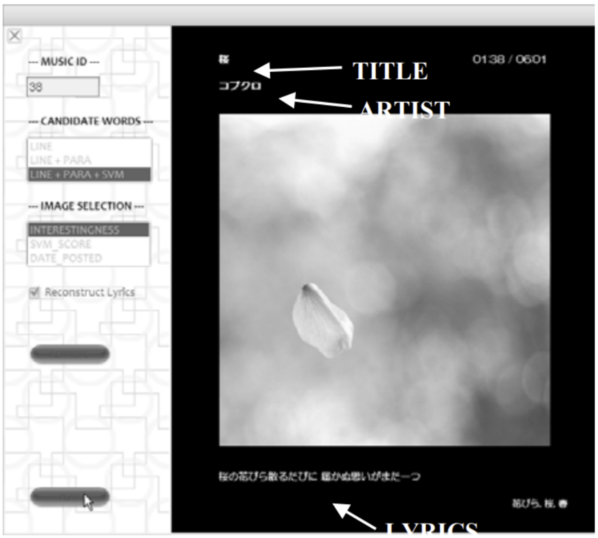
2.5.1 Music Slideshow

The first system is an Automated Music Slideshow Generator [13], which displays images based on each line of the lyrics, as seen in Figure 2.3. There are three distinct stages to this system, which it outlines as Candidate Image Retrieval, Image Selection, and Synchronized Playback. For the first stage, they collect a number of images that have tags corresponding to the set of nouns in a given line. Additionally, for a line that has no nouns, the image that fills that space is based on the general impression of the song; this impression is determined by running the set of all nouns in the lyrics through a text-based classifier, trained on a database of manually tagged words in relation to a Season, Weather, or Time.

This general impression is a factor in the next step, Image Selection, the step in which the image to be displayed is chosen from the candidate images for each line. This is done by calculating the impression score using the specific tag and the general impression of the lyrics; this is posited to increase the relevance of the image tags and decrease the prevalence of popular tags in the image selection process. The image that has the highest impression score for each image is selected for display.

The third step is Synchronised Playback, which involves referring to known lyric information for the song and making adjustments to these timings based on the length of the line in proportion to the length of the song.

This previous system differs from the system outlined in this dissertation in a number of ways. The first and largest difference from the previous system is that it assumes that line information is available for the given song, which is different from our system, as it has no line information prior so must generate or fetch its own line information for a given audio. Secondly, the previous system shows images that are influenced by multiple factors and are even shown in blank sections, whereas this system searched for images without further context, and even filters out potential extrinsic values. Thirdly, the previous system searches for a single word per line and attempts to estimate changes in timings based on length, although this method is fairly similar to the Lyric Analysis and Caption Extrapolation methods in the inferring timings, the outline system can provide word accurate timings using the Forced Alignment method, which then has the timings padded similar to the previous system.



--- MUSIC ID ---
38

--- CANDIDATE WORDS ---
LINE
LINE + PARA
LINE + PARA + SVM

--- IMAGE SELECTION ---
INTERESTINGNESS
SVM_SCORE
DATE_POSTED

☑ Reconstruct Lyrics

桜
コブクロ
0138 / 0601
桜の花びら散るたびに 届かぬ思いがまた一つ
花びら, 桜, 春

Figure 2.3: *This figure shows the user interface for the Automated Music Slideshow Generator [13].*

2.5.2 Visualyre

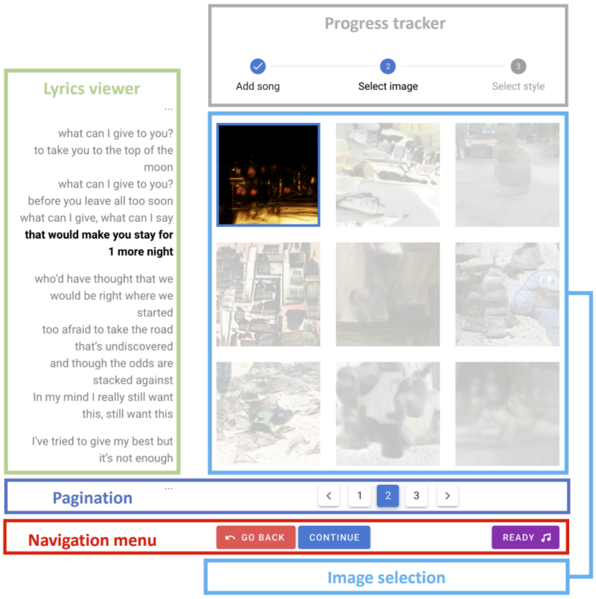
The second system is called Visualyre [], which is a web based application in which an image is generated based on the semantics and mood of song's lyrics; such a system was created for musicians to easily generate abstract album covers which are cohesive with their music. The various modules of this system includes text-to-image generation from lyrics, audio analysis to determine the mood of the song, and Style Transfer which modifies the generated images to align with a collection of mood images. The process is facilitated through a user interface with input required from the user to select a generated image, as seen in Figure 2.4.

The first stage is image generation, which is generally done through Generative Adversarial Network (GAN) models which are comprised of two neural networks, a generator attempting to recreate images from a dataset and a discriminator that attempts to differentiate between images that are real or generated. The training of the model is complete once generated images cannot be reliably differentiated from real images. For this system, a Dynamic Memory Generative Adversarial Network (DM-GAN) model is used to refine and conserve quality from images in past cycles of training. During the system processing, an image is generated for each lyric line and is selected by the user to apply the style upon.

For audio analysis, a binary classifier is used to classify the mood of the audio into four different moods: anger, happiness, sadness, and relaxation. Each of these categories have a corresponding set of 'emotional' images for the next stage of Style Transfer prior to system processing.

Style Transfer refers to changing the style of an image to match the style of another without loss of characteristics. The basis of this is in deep-learning through finding the commonality in neural representations between the two prior images to form a third image with impressions of both previous images. The results of the audio analysis are normalised to produce a Mood Probability which increases the chances that a mood will be sampled from the collection of that mood, for combination. A total of nine stylised images are produced for the user to select for download.

This is similar to our system in that it creates a product of increased modality from the textual and audio components.



what can I give to you?
to take you to the top of the moon
what can I give to you?
before you leave all too soon
what can I give, what can I say
that would make you stay for 1 more night

who'd have thought that we would be right where we started
too afraid to take the road that's undiscovered
and though the odds are stacked against
In my mind I really still want this, still want this

I've tried to give my best but it's not enough

Figure 2.4: This figure shows the user interface for the Visualyre system [2].

2.6 State of Lyrical Transcription and Alignment

A system capable of automatic generation of time-aligned lyrics has applications which include the creation of karaoke scrollers, text-based song retrieval, and automatic censoring of explicit content. The earliest attempts to adapt speech recognition technologies would ultimately fail due to the difference in assumptions made by the systems between the structure of speech and music. For speech recognition models there is an assumption that speech and noise are statistically independent, however speech in music is often highly correlated with various musical components, regarded in this case as noise [3]. Another assumption in speech recognition models is that there is often phonetic consistency which contributes to the intelligibility of the spoken word, however the variations in the frequency ranges, timbre, tempo and dynamics need not lend itself to intelligible lyrics within the song [43].

Pre-trained recognition models (or phoneme detectors) were initially used in hopes to bypass issues in having little accurately labelled music recordings in which to train from at the time [31, 21]. However, these would suffer from low accuracy due to previously stated issues, such as variable syllable duration and low word intelligibility [43].

Due to the background music introducing overpowering noise there were systems created which were adapted to work for solo-singing tracks which could recognise lyrics to a fair degree of accuracy [21]. However, it is very rare for songs to be solo-singing tracks; the most common form is polyphonic music, which have multiple musical components. Therefore, attempts were made to create systems that are capable of performing voice-separation techniques to extract this clean audio [30]. However, this would often leave artifacts behind which would leave the phonemes unrecognisable [9].

Beyond these initial attempts, there have been multiple different techniques utilised to improve the accuracy of lyric recognition. One such approach was to apply various processing methods to the music [12]. Other previous systems made assumptions about the input values to simplify the process, such as assuming each line or phrase had already been pre-aligned [31]. Whilst other systems attempted to recognise and impose alignment upon similar repeated phonetic sequences [47] or assumed that musical chords were recorded in the line information [43]. Additionally, these systems depended upon complicated training and parameter optimisation processes, which could negatively impact on the duration of training or the reliability of the end system [43]. These systems with their multiple complex steps created pipelines that, if not jointed optimally, allowed for the propagation of mistakes throughout, corrupting the results [43]. Many of these issues persist today, even in many of the state-of-the-art systems which require extensive training with large datasets, whilst still providing low accuracy results [33]

2.7 Techniques of Keyword Extraction

For this system, we need to find meaningful words within the text of the lyrics transcript of a song, to be represented by images. These meaningful words are often called keywords, and there are various automatic keyword extraction methods available.

With a constant, and increasing in size, stream of documents being uploaded to the Internet, the majority of which have little to no tags pertaining to their content [4]. Due to the volume being uploaded in this state, there is no possible way in which they can be manually tagged, which has created the need for automatic keyword extractors to create appropriate tags for these documents.

2.7.1 Keyword Extraction Technologies

The applications for keyword extraction include summarisation, information visualisation, opinion mining, categorisation, and indexing [34], among others. This provides usual tools for journalists, librarians, and historians in which they can sort and search documents in the ever expanding number of available documents. With the number of documents exceeding 1 billion [34] it is no longer viable to assign keywords by hand. However, manual keyword annotation is still far more accurate than any state-of-the-art system to date [34], such that keywords generated by human annotators are differentiated as golden-words, or ground-truth words, in which is it generally accepted that these convey the true nature of a document [34].

There are two different classifications of keyword extractor, these being supervised and unsupervised approaches. Supervised approaches are the most popular method of keyword extraction, which makes use of discriminant features and machine learning algorithms which are trained to determine what is and is not a keyword [34]. This training process requires a lot of time as well as a large manually annotated corpus to learn the characteristics of the text. On the other hand, unsupervised approaches do not require training, and are therefore characterised by their independence of any particular corpus [34].

Unsupervised methods have various methods of analysis which encompasses statistical methods to graph-based approaches. An early example of a simple statistical method would be TF.IDF [26] which extracts keywords based on their term frequency; this is simple to implement, however it requires a large corpus which may not be obtainable. Another statistical method which relies on TF.IDF is KP-Miner [10], which uses TP.IDF and two additional factors of word length and word positioning in the document, to determine keywords. Finally, the statistical method that has the most in common with YAKE [34], is RAKE [42]. This method treats a document as a sequence of candidate keywords, delimited by stopwords, on which a matrix of co-occurrences of a term are stored. The term score, which determines if the word or phrase is key, is based on term frequency, term degree (number of terms that occur with the word), and the ratio between the term frequency and degree [34].

For the graph-based approaches for representing a text document, the most widely known model is likely TextRank [32], which represents terms as nodes with edges between nodes representing the degree to which these terms co-occur. A ranking algorithm is then used to sort nodes in decreasing order. Other models such as SingleRank [46] and ExpandRank [46] are variations of the TextRank model, with ExpandRank using the terms of k-nearest neighbouring documents to refine the results.

The earliest incarnation of a supervised model was called GenEx [45] which used a custom algorithm which could outperform generic machine learning algorithms, such as decision trees. The most well known of supervised models would be KEA [48], which determined keywords using the Naïve Bayes machine learning algorithm and the additional factors of TP.IDF and the term's first occurrence[34].

2.7.2 YAKE! : Yet Another Keyword Extractor

YAKE! is an open-source unsupervised keyword extraction method, which can analyse a single document, scalable to any size, without need of a training corpus; the algorithm uses text statistical features to determine the probability of a word or phrase being keyword or key phrase, independent of the term's frequency; this is in contrast to other methods which provide a binary classification of each keyword. The YAKE! system is also able to extract key phrases, which contain interior stopwords better than other state-of-the-art techniques [34]. Examples of interior stopwords in Charles Dickens's novel, *A Tale of Two Cities*, are the words "of" and "a". Due to these factors, YAKE! is considered to be corpus, language and domain independent.

The efficacy of YAKE! was compared 11 models over a dataset which included 5 different collections, each in a different language: English, French, Spanish, Portuguese and Polish. Of these models, 10 are unsupervised models, which include TF.IDF, KP-Miner, RAKE, TextRank, SingleRank, and ExpandRank, among others, and 1 is a supervised model, this being KEA. YAKE! outperformed all unsupervised models and was able to match or surpass the efficacy of the state-of-the-art supervised model [34]. The adaptability of the system to any text without prior training, made it an attractive choice to be applied for the extraction of key words and phrases from song lyrics.

3 | System Requirements and Design

The chapter outlines the purpose of various components of the system in reference to the dissertation briefing, as well as the justifications for the design decisions made for each. To highlight the priorities of the project the MoSCoW method [38] will be used in this section to state the importance of the features implemented by each component. This method is one of the practices of agile project management, usually used by teams of developers to focus their attention on the most pressing component in that stage of development. The method is broken down into four different sections: *Must have*, *Should have*, *Could have*, and *Won't have*. During this section, different features will be placed in either of these four categories, with justifications being in reference to the dissertation brief and what is currently possible to implement. Additionally, all *"Must have"* statements, or variations of, will be the direct result of the dissertation brief, with all other statements reflecting the degree to which they further these main directives. These statements shall be made in the *system requirements*, *design* sections below, as well as in *Chapter 5* during discussions of *Future Work*.

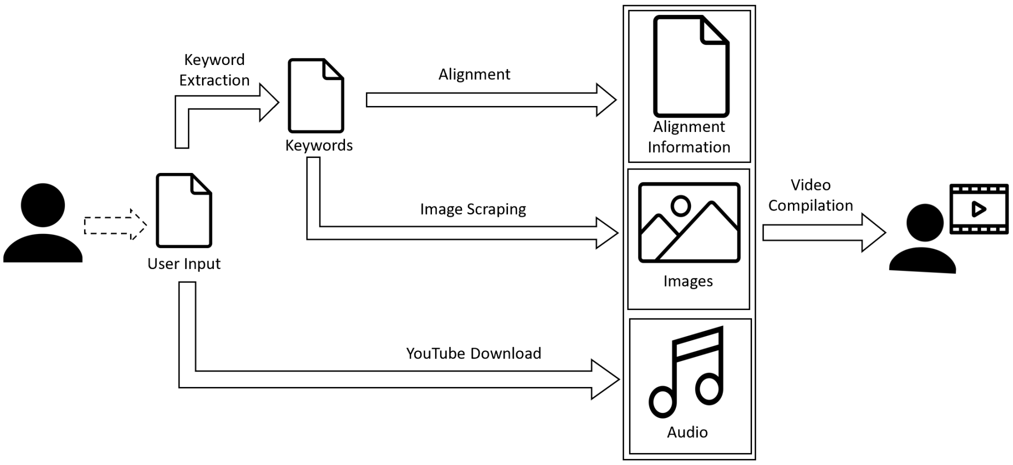


Figure 3.1: *A high-level overview showing the steps in which the system converts user input into the output video.*

3.1 YouTube Downloader

The YouTube download component is responsible for fetching the video or audio of a user-selected YouTube video. The determination of either audio or video is dependent upon the alignment method that has also been selected by the user.

3.1.1 System Requirements

The system *must have* the ability to allow the user to indicate which source YouTube video should be processed. It *must also* have the ability to extract the video or audio files from the selected source video, and it *should be* able to extract the audio of a video if it should need to do so. This process *should be* based on a reliable method of retrieval.

3.1.2 Design

The system *should be* able to identify source video through the YouTube hyperlink provided through user input. In the case that the user has selected to use either the *Forced Alignment* or *Caption Extraction* method the system *should be* able to extract the audio from the source video. However, in the case that the user has selected to use the *Lyric Analysis* method the system *should be* able to download the source video locally, and from this downloaded video it *should be* able to isolate the audio from the video and extract frames periodically from the source video for later use.

3.2 Keyword Extraction

This component is responsible for selecting words which are to be examined during the lyric alignment component.

3.2.1 System Requirements

The system *must be* able to retrieve the transcript of a song and the system *must be* able to identify a number keywords from within this textual element. It also *should be* able to identify that n-gram phrases are key phrases, where appropriate.

3.2.2 Design

In the case that the user has selected to use either the *Forced Alignment* or *Lyric Analysis* method, the system *should be* able to fetch the transcript of that song from a song database, using the provided information from the user. However, in the case that the user has selected to use the *Caption Extraction* method, the system *should be* able to fetch the captions for the YouTube video, if they exist, and *should be* able to make available the timing information within the captions to the Lyric Alignment component.

The system *should be* able to handle any song recognised by the song database given the provided information identifies a transcript which corresponds to the source video. The assumption that the YouTube video has accurate source material applies here. This would allow covers, which are characterised as being faithful to the lyrics of the original but with a different artist, to work with the system. Although, this would exclude remixes of songs, which are characterised as having multiple additional elements that the original song(s) do not have, unless the remix is considered as a song in its own right within the song database or has available faithful captions for the *Caption Extraction* method.

The system *won't be* able to reliably determine the artist name and song title from the available metadata of the video, which is an issue discussed previously in section 2.3.2. Therefore, the system *should be* able to take these values as input from the user.

An assumption made about the input song is that it is exclusively in English. Therefore, the system *won't be* able to handle videos which are not in English.

3.3 Image Scraping

This component is responsible for fetching a number of candidate images to represent each key word or key phrase which has previously been identified.

3.3.1 System Requirements

This component *should be* able to interface with the *Keyword Extraction*, described in section 3.2, to fetch the keywords for a given song. The system *must be* able to retrieve a set of candidate images for all identified key words. This process *should be* reliable and scalable to any number of keywords.

3.3.2 Design

The system *should be* able filter the output of the search results so as to remove pictures that are not appropriate imagery. However, it *will not* be able to verify the contents of the image, instead it must rely on the tags provided by the images.

3.4 Lyrical Alignment

The lyrical alignment component is responsible for analysing the given YouTube video through various methods to detect occurrence of keywords within.

3.4.1 System Requirements

The system *must have* the ability to detect the presence of keywords within the piece of music, and *must be* able to determine when said keywords are spoken. This determination *should be* as accurate as possible and the system *should have* both a reliable and scalable method to retrieve these timings.

3.4.2 Design

As outlined in chapter 2 in table 2.1, there are various subtypes of videos each with different characteristics. Therefore, we propose a total of three different methods to exploit these various characteristics: the *Forced Alignment* method uses a lyrical alignment model to map known lyrics to an audio file, working on all subtypes of music videos from ‘still’ to ‘classic music videos’; the *Caption Extraction* method uses the in-built closed caption accessibility feature in YouTube [16] which contains timings information for each phrase of the captioning; and the *Caption Extraction* method uses image-to-text models to extract the lyrics within the frames of ‘lyric’ and ‘embedded lyric’ videos to determine which phrases contain keywords. Whilst the *Forced Alignment* is accurate to the word, *Caption Extraction* and *Lyric Analysis* are both accurate to the given phrase, making them similar to the previous technology outlined in section 2.5.1. However, unlike the technology in section 2.5.1 this system is able to generate timings information for any song without prior information gathering.

Despite having the three different methods, the system *won’t have* the ability to determine the appropriate method with which to treat the video automatically. The system instead will rely upon the user to select whichever method for the video is most appropriate, which assumes that the user is able to determine what is appropriate. Even if this assumption is wrong the system *should be* robust enough to recover from incorrect input: the *Forced Alignment* method *should* always work for any given music video; in contrast *Caption Extraction should* always fail in the instance of the video having no captions, which is as intended for that method is impossible for such a video; however, if the *Lyrics Analysis* method is used with an incompatible video subtype this will result in the retrieval of no timing information. This is the only sense in which the assumption to trust user input fails in this instance. However, given the next assumption is not broken then this point is moot.

The additional assumption is that the YouTube video has accurate source material. This means that the lyrics that are being detected for *Lyrics Analysis* method and that the captions extracted by the *Caption Extraction* method, both correspond to the words sung. These two methods do not use audio as a factor, but merely infer timings through a source which itself should correspond directly to the audio. This is a reasonable assumption to make, as captions and lyrics are both characterised by their faithful portrayal of the source material.

Finally, the output of all lyrical alignment methods *should be* a generic dictionary storing the keyword and the corresponding duration of display.

3.5 Video Compilation

This component is responsible for combining the products of previous interfaced components to generate the final video output. This culmination of the final video is *the* aim of the system outlined in the dissertation briefing.

3.5.1 System Requirements

The system *must be* able to produce a video, with identical audio to that of the source video and imagery relating to the keywords found within the lyrics. It *must be* able to retrieve the audio, timings, and images from the various components where appropriate. This process *should be* reliable and scalable to any number of images that are required to be displayed.

3.5.2 Design

From the retrieved timing information, the system *should be* able to select an image which corresponds to the keyword associated with each timing entry. The system *won't be* able to verify the content of the video, as stated in section 3.3.2, therefore the system *should* select an image at random, from the retrieved the set of candidate images; this component relies on the prior image retrieval with its filters for the accuracy of the available images.

4 Implementation

The following section outlines the various components of the system as well as the choices made during the implementation of the components. This shall lead into an explanation of how to install and operate the system, as well as demonstrate the output of the system.

Figure 4.1 shows the entirety of the system, showing the various components and how they interact. Not all methods use the same components as is denoted by the coloured arrows. A coloured arrow as input/output for a component can be read as 'if this method is activated then this component will take/produce this input/output'. A black arrow as input/output denotes all methods and can be read as 'this input/output is always required/produced'.

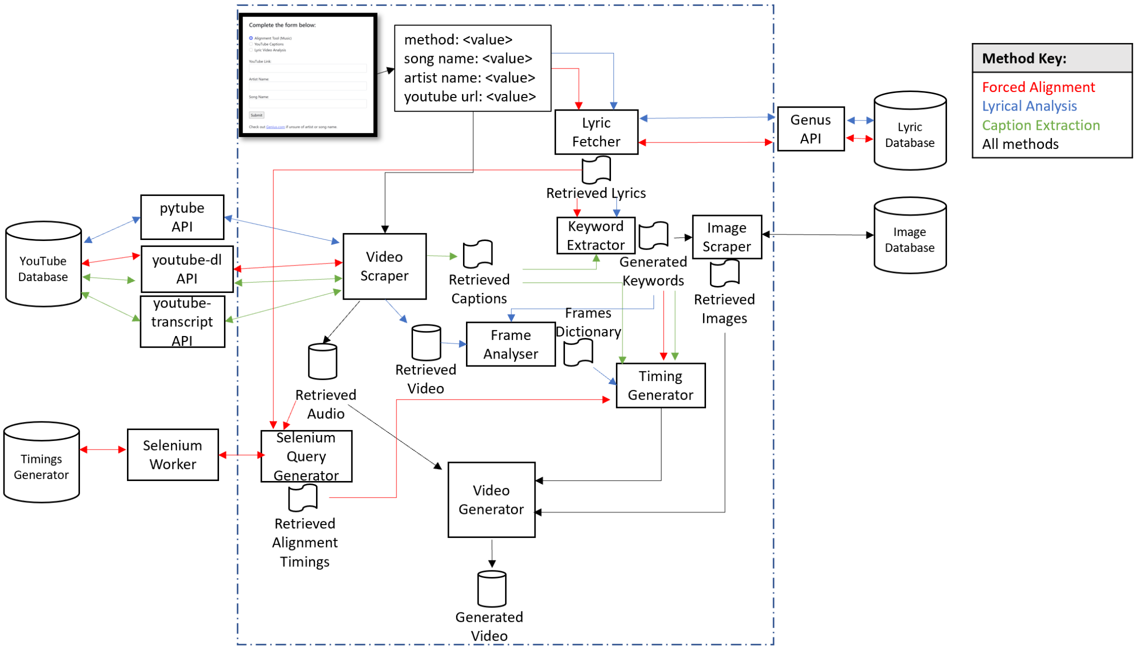


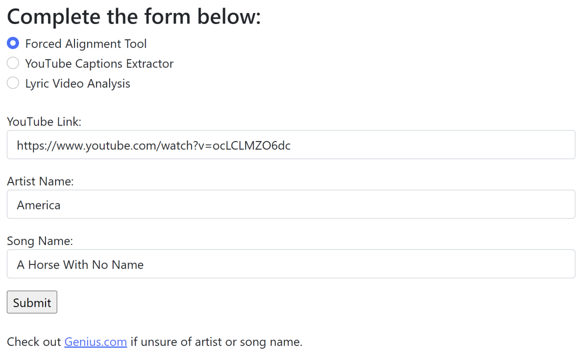
Figure 4.1: Figure demonstrating the various steps taken by the system and interaction within and without the system. The flow of of operations depends on the method of alignment, as denoted in various colours, that is selected by the user.

4.1 System Components

4.1.1 Front-end Technologies

As a way for the user to interface with the system the high-level python web framework, Django[7], is used. Various webpages are available within the system, such as the instructions to run the system, the collection of videos(which is in-built and extendable by the user), and the homepage containing the form which allows for the interaction with the back-end of the system. For reference of webpage appearance refer to figure 4.2. Django was chosen as the system is written entirely in python this allowed for seamless integration of front-end and back-end systems, making for easy use and deployment of the system.

The user is required to fill out the contents of the form, which includes: the method of alignment, YouTube link, artist name, and the song title. If the user is unsure of these details of the song then the user is prompted to visit the Genius website [15] from which the can user search for the song within the YouTube video. Upon submission of this form, the user will be placed in a loading screen, and shall be redirected once the video has been compiled.



Complete the form below:

● Forced Alignment Tool
○ YouTube Captions Extractor
○ Lyric Video Analysis

YouTube Link:
https://www.youtube.com/watch?v=ocLCLMZO6dc

Artist Name:
America

Song Name:
A Horse With No Name

Check out Genius.com if unsure of artist or song name.

Figure 4.2: Figure displaying the form which must be filled in my users to begin video generation.

4.1.2 YouTube Downloader

Once the user has submitted the form, the method of alignment and YouTube link are provided to the Video Scraper component, which is responsible for gathering all content of the video corresponding to the YouTube link. In the case that the method is Lyrical Analysis, the pytube[41] python package takes a YouTube ID for a given video and returns the video. Additionally, the OpenCV[24] python package is used to extract the audio for later use. In the case that the method is either Caption Extraction or Forced Alignment, the youtubedl[14] python package is used to extract only the audio of the video.

This component is the most volatile and unfortunately there appears to be no mitigating factors that can be done to change that fact. The issue lies in YouTube taking preventative measures against what they refer to as *'stream-ripping'* through constant technical improvements[6]. Due to this, packages that are able to access YouTube content are also caught in the cross-fire, and they become unable to access the content. This problem will be discussed further alongside installation guidance in section 4.2.

4.1.3 Keyword Extraction

Lyric Fetcher The prerequisite to the function of the Keyword Extraction component is that it has a lyric transcript to act upon. This is generated by different components depending on the alignment method which is being utilised: for Forced Alignment and Lyrical Analysis, the transcript is retrieved from a song database by the Lyric Fetcher, and if the method is Caption Extraction, the Video Scraper component is used to construct the transcript from collected caption information.

The Lyric Fetcher uses the Genius python API [44] to access the contents of the vast collection of song lyrics held in the Genius database. To query the database this requires the song title and artist name, which has been provided by the user in the submitted form. The Video Scraper component, which is responsible for gathering information pertaining to the YouTube video, uses the python package, youtube-transcript-api [5], which returns a JSON dictionary containing the text segments within the captions. These text segments are compiled together and processed to remove non-ACSII characters, to form a whole transcript for the YouTube video. The captions information is retained and made available to other components for later use. This lyric transcript is then made available to all other components, such as the Keyword Extraction component.

The process of Keyword Extraction When the lyric transcript is generated it is broken down into segments, which are of an arbitrary size determined by the formatting of the lyrics in the Genius database, or of the length of each caption phrase. We employ the YAKE (Yet Another Keyword Extractor) tool [34], which has been discussed previously in section 2.7.2, to attempt to extract a total of 2 keywords or key phrases, in the instance of a bi-gram, which are collected in a set. Once all segments have been examined this collection of keywords are made available to the system for use in other components.

4.1.4 Image Scraping

This component is used by all methods of alignment as they all require images in which to represent the keywords of the video being processed. The only prerequisite to this component is that there must be an available collection of keywords that it can query.

This process relies on two python packages, those being BeautifulSoup4 [40], a common HTML scraper, and Requests [39], a simple HTTP library for python. Firstly, Requests is used to generate a Google Images query for a given keyword, with using Google Images built-in features to filter images which refer to various forms of media. The HTML of this resultant Google Image result is examined and instances of image sources are extracted. These sources are then used by the Requests package to download a total of five images, which are stored within the system. This process is repeated for each keyword within the collection. The resultant collections of images are stored within the system, available to be accessed by other system components.

During testing of this system some images that had resolutions that contained odd numbers were discovered to be incompatible with the packages used in the Video Compilation component. This was fixed by adding the additional condition that all images downloaded must have an even resolution.

4.1.5 Alignment

As previously stated three are there alignment methods outlined by this system: Forced Alignment, Lyrical Analysis, and Caption Extraction. All of the alignment methods described generate a word:time-interval map, which we refer to as the timings list, within the Timing Generator component. This timings list is a sorted list of objects, each object is comprised of a keyword and the timing related information, specifically the start time and duration in which the candidate image is to be displayed. Figure 4.3 shows an example of the internal structure of some of the timings within the timings list.

The Forced Alignment method provides word-level alignment of a given lyrics text to a given polyphonic music audio file. The implementation is carried out by the Selenium Query Generator component using Selenium [25] to access a web-based forced alignment model [22] in which the lyric transcript text and audio MP3 files are both submitted. The system awaits the return of the word-level accurate dictionary for each word, which is then converted into the generic timings list within the Timing Generator component and made available. An additional buffer is added to the timings returned to elongate the display time of the images that would otherwise be too brief.

Lyrical Analysis provides phrase-level alignment through the detection of the presence of keywords in frames. Firstly, within the Frame Analysis component we use the Opencv [24] tool to determine the frame rate of an MP4 video file from YouTube. We then use the frame rate information to extract a frame every 0.5 seconds, i.e. for a video of 28 frames per second (fps), every 14^{th} frame is saved as an image file (PNG). After extracting the frames, we analyse each to extract any textual information from them using the OCR tool, called pytesseract' [27]. If a sequence of frames have textual information containing the same keyword, then a timing in the timings list should be created for that keyword between the time of the first and last frame in which it is detected. For example, for a video 28fps if we find that the 14^{th}, 28^{th} and the 42^{nd} frames contain the same keywords, this would correspond to those keywords word being displayed from 0.5-1.5 seconds in the video. This timing information is then passed to the Timing Generator component so that the generic timing list can be constructed and made available.

Caption Extraction also provides phrase-level alignment through matching of keywords found in the text segments of entries in the stored caption information generated by the Video Scraper component. The Timing Generator component uses the caption information which also contains the start and duration of each corresponding text segment for a given YouTube video, to know the timings of the phrase containing the keywords. If more than one keyword is detected in a phrase then, the duration of that whole phrase is split amongst them, with the order of image display being determined by order of the keywords in the phrase. The timings list generated from the captions is then made available to the system.

4.1.6 Video Compilation

This component is the culmination of all other components. It takes as input the timings list, candidate keyword images, and YouTube video audio to render the output video through using the moviepy [49] package. For each timing in the timings list an ImageClip, an instance of the ImageClip class in moviepy characterised by the display of a single image for the duration, is generated. These clips contain a random candidate image corresponding to the keyword with their duration set to the duration of the image display within the timing. These ImageClips are composited together to form a VideoClip, an instance of the VideoClip class in moviepy which is the parent class to ImageClip. This composite displays images within the VideoClip corresponding to the timing of each image in the timings list. Finally, the song's audio is applied to this VideoClip, which is then rendered and the resultant video is saved in the file system. The process of video compilation is now completed and the user interface is redirected to the video which has been created.

4.2 How to Setup and Use the System

4.2.1 Setup Instuctions

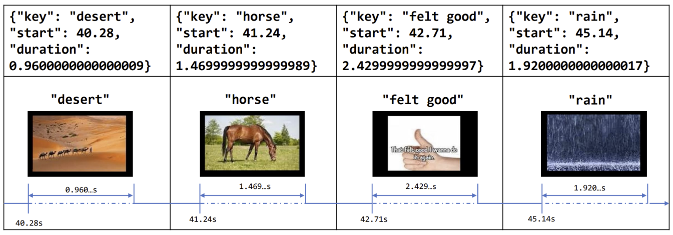
Instructions for how to setup the system can be found in the READ.ME within the source code folder. These instructions will include the step-by-step guide to the prerequisites of the system, as well as how to set up the python environment and the Django server within it.

System volatility As stated in section 4.1.2, the most volatile aspect of this system is the Video Downloader component which has fluctuating viability with changes made to the YouTube system.

4.2.2 Instructions for Use

The system, when initialised, will open onto the homepage in which the form to begin the compilation of the video will be found. The begin the process of compilation, operate the system as follows:

1. Select the method in which to process the video, by selecting the appropriate radio button. For example:
   ● Forced Alignment Tool
   ○ YouTube Caption Extractor
   ○ Lyric Video Analysis
2. User will be required to provide the following information:
   - YouTube Link
   - Name of the Artist
   - and Title of the Song
3. Once the details for the video have been filled, press the Submit button to begin compilation.
4. User will be redirected once processing has completed and the video will play.



```
{"key": "desert",
"start": 40.28,
"duration":
0.9600000000000009}
```

```
{"key": "horse",
"start": 41.24,
"duration":
1.4699999999999989}
```

```
{"key": "felt good",
"start": 42.71,
"duration":
2.4299999999999997}
```

```
{"key": "rain",
"start": 45.14,
"duration":
1.9200000000000017}
```

Figure 4.3: Figure revealing the internal timing dictionary from the processed videos, and the images which are displayed by the entries in that structure.

5 Evaluation

5.1 Product Survey

The purpose of this survey is to have a group of individuals provide valuable user feedback information on the videos produced by the system.

5.1.1 Survey Design

As this experiment involved human participants, it was subject to the ethical guidelines as outlined by the Department of Computer Science within the University of Glasgow [35]. This involved informing each participant about the nature of study and gathering their explicit consent, with verification of their understanding of the information they were given. To have a thorough distribution of individuals propositioned, the goal number of participants for this experiment was set to N=20.

The structure of the survey was as follows: individuals were asked to verify their understanding and consent to participate, as previously stated. They would then watch either a 'still' video, which was found on YouTube, or a video generated by the system; both of these videos were of the same song. The type of video given to the participants was programmatically determined in a random pair-wise manner, i.e. if a participant is randomly assigned a video type then another participant will receive the opposite. This forces an even distribution across the number of participants being exposed to either type of video whilst the partitioning of video types remains random. Each participant views a total of three videos, after each of which they will answer questions regarding the video they have just watched. If the video they watched was a 'still' video they would only be asked to rate their user experience. However, in the instance that the video was the corresponding generated video, then they would be prompted to rate not only their user experience, but also rate the perceived correlation between the images and the lyrics, and to rate the perceived accuracy of the alignment. Therefore, the individual may be asked a minimum of three or a maximum of nine ranked choice questions. Finally, they shall be prompted to provide their general impression of the videos and any insights or suggestions they may have for the produced video.

5.1.2 Survey Objectives

The user experience shall be quantified by the user and these values shall undergo quantitative analysis with reference to the NULL and Alternative hypotheses [11]. The NULL hypothesis, H_0, states that there exists no statistically significant relationship between variables. Whilst the Alternative hypothesis, H_a or H_1, posits that there does exist a statistically significant relationship between the variables. For this aspect of the experiment, the H_0 is:

> There will be no difference in user experience between that of the sampled still videos and the videos generated by the system.

Conversely, the H_a is :

> There is a difference in user experience between that of the sampled still videos and the videos generated by the system.

In order to reject H_0 in favour of Ha we must perform a p-test upon the retrieved user experience scores, and find that the calculated p values are less than 0.05. However, if the calculated p values are greater than 0.05 we cannot reject the H_0 as being true. Neither the NULL or Alternative hypotheses are to be considered true or false under any pretence, as this form of significance testing merely states the degree to which the data warrants favouring one hypothesis over the other.

For the analysis of the perceived correlation between the images and the lyrics and the perceived accuracy of the alignment, a different method must be used as these metrics do not apply to the 'still' video and therefore only apply to each of the generated videos. Therefore, instead of p-testing there will be an analysis of the tendency of users to agree upon the ranking for each video. Each video shall be generated through a different method of alignment; as the keyword extraction method remains constant and for any given alignment method the same keywords should be produced, therefore creating a comparative evaluation of alignment upon the perception of image relevance. Which method is used for each song can be found in table 5.1

Method	Song Name
Forced Alignment	A Horse With No Name
Captions Extraction	Roar
Lyrical Analysis	What Makes You Beautiful

Table 5.1: *Table stating the methods by which each video was generated by the system.*

5.1.3 Survey Results

For this experiment, the number of participants exceeded that of the goal number of participants, such that N=24. The results of the user experience ratings for each generated and generic video pairing are represented in figures 5.1, 5.2, and 5.3; to view these values in a tabular format refer to table A.1 in the appendix. The results of the timing and imagery questions are presented in figure 5.7 and 5.5 respectively, with the values for both represented in a tabular format in the appendices in A.2 and A.3. Figures 5.4 and 5.6 display the same values as in figure 5.7 and 5.5 respectively, in the box plot format to more easily represent the distribution of values.

Interpreting User Experience Rating As stated previously, these values from the user experience will be interpreted through p-testing to determine the significance of the values, which are represented in table 5.2. Referring to the p-values calculated in table 5.2, all values have exceeded the value of alpha = 0.05 so we can not hold to a degree of certainty that these results are significant. This means that from the data that has been collected we are unable to reject H_0 in favour of H_A.

Interpreting Timing Rating Figure 5.6 represents the distribution of votes in terms of timing rating for each of the given songs within a box plot. For the songs, "A Horse With No Name", "Roar", and "What Makes You Beautiful" the average timing score is 3.25, 3.083, and 2.528 respectively, and these are denoted by the crosses on each box for a given song. These results show that "What Makes You Beautiful" had the lowest average score and "A Horse With No Name" had the highest average score.

However, the comparative difference between Q1 point, which separates the lower 25% of the data from the rest, and Q3 point, which separates the upper 25% of the data from the rest, varies between the different song; this variance within the data is visualised by the size of the box for each of the box plots in figure 5.6. Examining the box plot for the song "Roar" we can see that the size of the box is comparatively smaller than that of the other songs, suggesting that participants are in greater agreement about the perceived alignment of the generated video, which appears to be slightly positive. In contrast, for the songs, "A Horse With No Name" and "What Makes You Beautiful", have a much larger box which indicates that the question elicits far more varied answers from the participants.

Interpreting Imagery Rating Figure 5.4 represents the distribution of votes in terms of imagery rating for each of the given songs within a box plot. For the songs, "A Horse With No Name", "Roar", and "What Makes You Beautiful" the average imagery score is 3.417, 3.583, and 2.385 respectively; these are denoted by the crosses on each box for a given song. Therefore, "What Makes You Beautiful" had the lowest average score and "Roar" had the highest average score.

Examining the box plot, within figure 5.4, for the song "What Makes You Beautiful" we can see that the size of the box is comparatively smaller than that of the other songs, suggesting that participants are in greater agreement about the perceived alignment of the generated video, which appears to be slightly negative. In contrast, for the songs, "A Horse With No Name" and "Roar", have a much larger box which indicates that there is not the question elicits far more varied answers. However, despite being variable, the answers appear more neural than positive.

5.1.4 Interpreting Open Questions

This section will summarise the feedback which was received from participants after they had watched their allocated videos. This summary shall inform a commentary on the impression that the individuals had of the product and the applications of their insights into improvements that could be made to the system to increase overall enjoyment of the product. It should be noted, that of the individuals that were asked this open question, there were 5 that otherwise indicated that they had no comments further than of their responses in the ranked choice questions.

The Concept of Automatic Videography One individual made a comment which challenged the premise of the system, stating that: “The pictures representing random lyrics makes [for a] bad viewing [experience] for me”; this comment being in reference to the premise of the system itself. We can speculate that as they made this comment in reference to their own enjoyment, they were not suggesting that the product was inherently flawed, merely that they themselves would not enjoy such media. As not all forms of media need be enjoyed by everyone, the availability of such a tool to create these forms of media do not need to have a negative effect upon the media landscape. However, this form of bias against the premise itself may have influenced the individual’s rankings for the videos, of which they scored the lowest possible rating for all ranked questions except two.

Lyric Semantics A number of participants elaborated upon their rankings which were given in the previous sections. Some noted that in their opinion of the images were overall accurate to the word of the lyrics, whilst others noted that they had penalised the video for relevancy in the instance that the image was semantically different to the sentence. This was highlighted by one participant, stating that although the semantic, and actual, phrasing of the song was “no pain”, which appears in "A Horse With No Name", the image that was shown instead corresponded to the keyword “pain”. Additionally, another participant noticed that the system had identified the keyword “champion”, which contextually referred to an individual, yet an image which referred to *Champion*, a company, was displayed. The former example presents a different paradigm for the imagery of the product, where the imagery represents the semantics, or meaning, of the song lyrics rather than the *word* of the song lyrics. Such an interpretation presents a viable expansion to the current system. The addition of a semantic component to determining images that should be displayed is similar to that of the system in section 2.5.1. The latter example refers to a failure in the image retrieval system, where the semantic filters in place to remove images, such as those referencing to various media and companies, are insufficient to remove irrelevant imagery.

Alignment and Blank Spaces A further elaboration made by participants was that of their impressions and expectation of the alignment of images. Some of this feedback was contradictory in places, such as one participant stated that the alignment of images was ‘unpleasant’, whilst another stated the alignment made it ‘quite funny to watch’. As these accounts are anecdotal, inferring much from these statements would be inappropriate, however this may warrant an exploration of the connection between personal preference and enjoyment of alignment. This also relates the biases within the design and implementation, as the timings of the forced alignment method are artificially elongated so that the images appear longer. This decision was made as during the implementation of the system, the display time was perceived to be too short, leaving large blank spaces, and this change was made to the system to counteract that. This user feedback contradicts the assumptions of that addition to the system, and therefore should be reconsidered.

Another common comment was that of the prevalence of blank spaces in the images. The existence of these blank spaces are by design as in these moments there would be no actionable keywords, and therefore no images. Therefore, during instrumental sections these sections should remain blank. Despite this, some candidates stated that having these blank sections negatively affected their experience of the video. One participant suggested that rather than leaving these spaces blank the system should instead fill these spaces with images that correlated with the song as a whole. It is a valid observation that these blank spaces are prevalent, however as the system’s goal is to represent the lyrics at a given moment, this may confuse users further; the constant contextual switching between images which are directly relevant to generally relevant to the lyrics would likely negatively affect the users’ perceptions of the relevance of all images. In contrast to these instrumental sections, prolonged sections consisting of non-lexical vocables, which are characterised as repetitive non-sense syllables [1], were pointed out by a participant to produce equally non-sense images. An example of this is the section in "A Horse With No Name" where the lyrics and vocals consist of “La La La” repeatedly, which as the participant says “[produces] pictures of random celebrities”. This is a failure of the keyword extraction method, as it is unable to recognise that, unlike most textual documents, the text contains both meaningful words and non-sense. However, a modification to the system to ensure that the keywords extracted are not nonsense may not improve the user experience overall, as removing those images would increase the prevalence of blank spaces in the generated video.

Image Quality and Other Representations Numerous participants also critiqued the images and the methods of their display in various ways. One of the larger aspects of user responses was the image quality being low, with one participant noting that the use of non-stock images ‘broke immersion’. Addressing the former, this is due to the Video Compilation component compensating for the diverse range of resolutions in selected image; the solution to this would involve the retrieval of images of the same, or similar, resolutions to lower this disparity. For the latter comment, it may be infeasible to have all images be stock images as we can speculate that not all concepts covered by songs are encapsulated within this form. However, this idea could be incorporated by having a preference within the system towards the retrieval or display of stock images. This would necessitate additional checks made upon the images, which we have already stated as a limitation of the system, but may be incorporated in future iterations.

Some participants suggested the addition of more images, meaning in the place of one image there should be multiple, whilst others suggested that image transitions could be used or even the replacement of images with videos instead. All of these suggestions have the theme of having additional dynamic components; the former two suggestions could be incorporated in the current system, however the latter suggestion would exponentially increase the complexity of the system. As previously stated in section 3, there are already assumptions involved in the image retrieval process, as well as the random element introduced into image choice during video compilation in lieu of an image verification component. In the instance these images were to be changed into videos, these assumptions will be strained as, in the worst case, every frame of a video would need to be verified to ensure relevant imagery, and perhaps more importantly, ensure no explicit content is displayed. However, as previously stated in section 2.3.1, there currently exists no system which is on par with that of a human reviewer. Without such a system, this may require implementations of displaying videos to be done through access to a collection of videos, which have been reviewed by a human and could be used to represent moods or the semantics of the lyrics.

The Music Some participants stated that the song choice influenced their decision for their given ranking. We can speculate on two ways in which this could be mitigated: having participants listen to a single song four different times, one for the 'still' and three for all methods of alignment, or have a larger collection of songs for the user to evaluate. Having the users be exposed to a large collection of music would be preferable for information collection, however, under ethical guidelines, considerations must be made to minimise how strenuous the process is upon the user. We could speculate that exposure to a single song may also be perceived as a monotonous task for the users such that, even if the sequence was random, later videos in this sequence would suffer a reduction in user experience. A possible mitigating factor may be to play multiple songs from the same artist; to have music with a similar sound, yet variable enough to not be monotonous. This would also help isolate the confounding variable of the 'hater bump' as described in section 2.2.2, as the artist would remain a constant.

This experiment follows more closely with between-group testing, i.e. having different people testing different conditions, rather than within-group testing, i.e. having the same people testing the same conditions. This allowed the experiences of the 'still' video to act as a control towards that of the generated videos. However, in the instance that the viewed video provoked an extreme response within the individual, such as that in line with the 'hater bump' phenomenon, that may not be represented in the reported experiences of the corresponding video. These responses may be balanced through the randomisation process of video exposure that was already used, however more consideration is needed to detect these outlier data points and the mitigation through other means.

5.1.5 Survey Conclusions

In this evaluation we were unable to produce statistically significant results such that we were unable to reject H_0 in favour of H_A. Therefore, additional research is required to determine if this form of media does improve the user experience.

In terms of timing, the video "Roar" which, as noted in table 5.1, was generated through the Caption Extraction method yet received more consistent and slightly more positive timing scores than that of the other songs. The method by which this video was generated is noteworthy, as the song "What Makes You Beautiful" was generated through Lyrical Analysis, which should provide similar phrase alignment to that of the Caption Extraction Method, yet was less consistent with participant scoring. Additionally, the song "A Horse With No Name", generated through Force Alignment, which is a word accurate method with an additional buffer, as stated in section 4.1.5, provide the best fitting alignment method, although this was not uniformly identified by the users. We can speculate from this data and from user testimony that this variability in alignment scoring may derive from blank spaces in the videos correlating with a perceived failure in alignment from the user. This aligns with user feedback that the song "What Makes You Beautiful" had far more blank spaces than that of the other videos comparatively and the inference that forced alignment should be the best fitted, which would cause longer blank spaces, yet it is perceived not to be. As this is speculation, future work should investigate this perception further.

In terms of imagery, the song "What Makes You Beautiful" received the most consistent and slightly more negative imagery scores than that of the other songs. From user feedback, an important factor in how a user rates an image is the semantics of the lyrics. We can speculate that this consistent negativity for the imagery in "What Makes You Beautiful" is due to the semantics of the lyrics and words themselves aligning rarely. The other songs were on average perceived better however had far more variable voting patterns than that of "What Makes You Beautiful". Extending the previous inference we could speculate that the semantics of the imagery in the other songs aligned more frequently as suggested by the higher average score, however the variability may be sourced from the perceptions upon whether the semantics or wording are being represented by an image. However, these inferences require further investigation before being accepted as true.

5.2 Future Work

5.2.1 Future Research

A possible future avenue for research would be into the effect that semantics or the word of the lyrics has on the user experience. The idea of meaning over definition relates to the idea of interpretation rather than translation [20], in relation to accessibility, therefore there may be some interaction between these topics.

As noted previously, investigating the user experience by applying various alignment methods to songs by the same artist could provide greater insights; allowing for the presentation of various songs to users to act as a mitigating factor against negative impressions of artists, whilst also exploring the alignment methods. It would also be interesting to investigate how sensitive or tolerant individuals are to disparity in alignment or to that of filling empty spaces in the video with general impression images.

Additionally, in the instance that image transitions were incorporated into the output video this may create the perception of blurring the lines between alignment which may be worth investigating. For example, the variable velocity of the image may allow the user to infer the duration of an image's representation, or allow the user to rationalise poor alignment as a product of the transitions themselves. However, this may depend on the identity of the song and, as discussed in section 2.2.2, users are less tolerant of changes in terms of video identity [29].

Not only should new lyrical alignment methods be explored as they develop, but lyrical transcription should also be looked into. Where lyrical alignment requires the audio and lyrics, lyrical transcription would only require the audio. Current lyrical transcription methods are far less reliable than the current best lyrical alignment models, as it is unguided by known lyrics; such a change to a lyrical transcription model would therefore reduce the amount of information required for a given song.

5.2.2 Modifications to the System

The following section will outline features that *should be* or *could be* added to future incarnations of this system. The *should* statements are modifications that when implemented would improve the system in its current state, whilst the *could* statements denote modifications that would improve the system, and would also change the nature of the system, but may be implemented as an option available to users.

Starting from the simplest modifications that *should be* implemented, the first is that of the removal of non-lexical vocables from any keywords. The system is already under the assumption that the songs are in English, therefore the removal of a combination of any non-English word or manually assigned common non-lexical vocables would be sufficient. The next modification that *should be* implemented is that of the addition of more filters for the removal of branded items, and this requirement can be combined with the condition that the system *should have* a preference for stock images where available, by examining the metadata of various images. The problem with metadata is less significant in this respect, as there is a plethora of images to use in this case, so the system can be more selective with images that have insufficient metadata. Another, simple change that *should be* made is the removal of the buffer which was instilled into the forced alignment method, as this is based on the false assumptions of the user enjoyment; in doing so that forced alignment method should display images only for the duration of that word in the lyrics.

Additionally, there *should be* the option for the users to validate images to ensure their contents align with that of the intended search, beyond that of the images' available tags. This would mean removing an assumption made in the system about the nature of the content.

The largest change that *should be* made to the system is that of having the forced alignment model run locally. As when the system makes a call to the web-based model, the system occasionally fails, although this would not be the case if the model was run locally; the advantage of doing so keeps this system light-weight. However, when these errors occur, the call must be made again, extending the time it takes for system processing.

The system *could be* changed such that the semantics of the lyrics are represented rather than the word of the lyrics. As previously stated, this is a different, yet equally viable interpretation of this same system. The user feedback suggests that this may also improve the user experience, therefore an option could be added to encompass both paradigms. The next modification is results orientated, in that there *could be* multiple implementations of the system that reduces the amount of blank spaces in the produced video. Such implementations *could be* to add textual lyrics in the blank spaces, add general song-related imagery in these spaces, or perhaps use a form of Style Transfer, which is one of the processes outlined in section 2.5.2 to combine the imagery of multiple images of a line; such implementations are numerous and the resultant reduction in blank spaces is desirable, as evidenced by the user feedback.

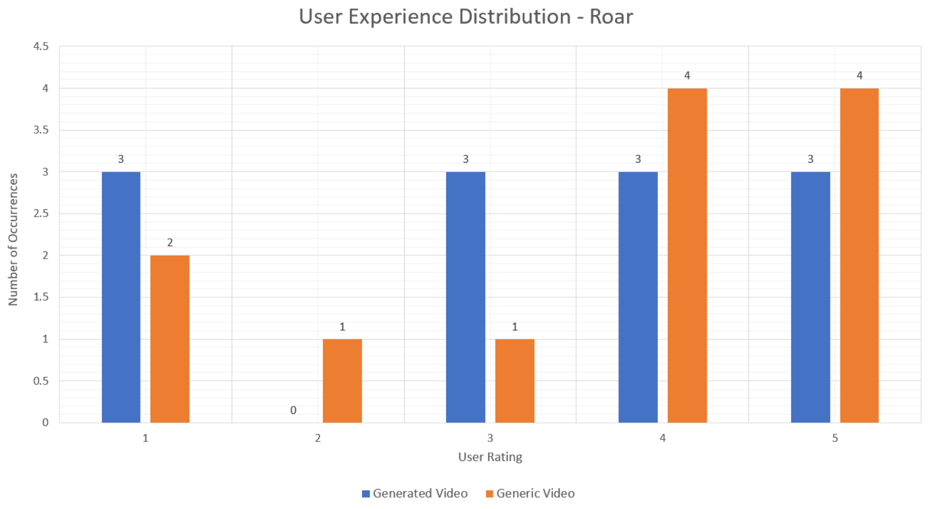


Figure 5.1: *The results of the user experience question referring to the generated video of Katy Perry's song,* "Roar".

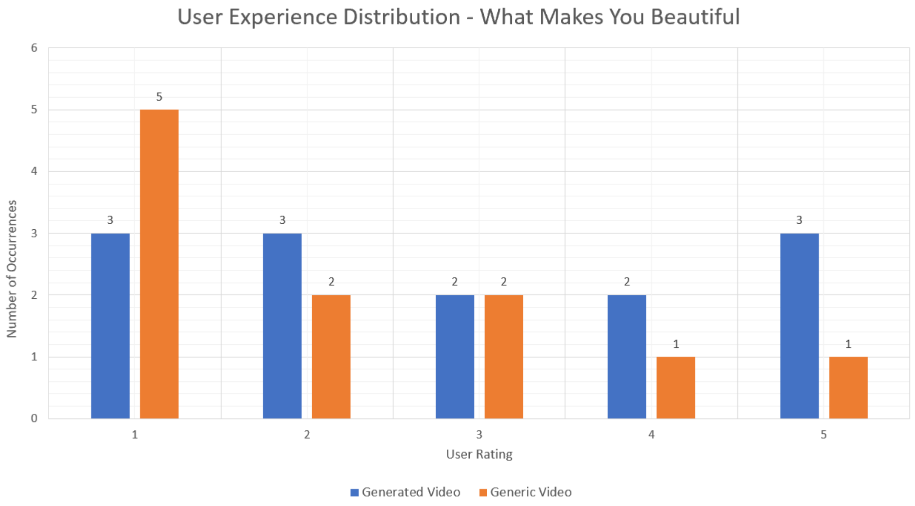


Figure 5.2: *The results of the user experience question referring to the generated video of One Direction's song,* "What Makes You Beautiful".

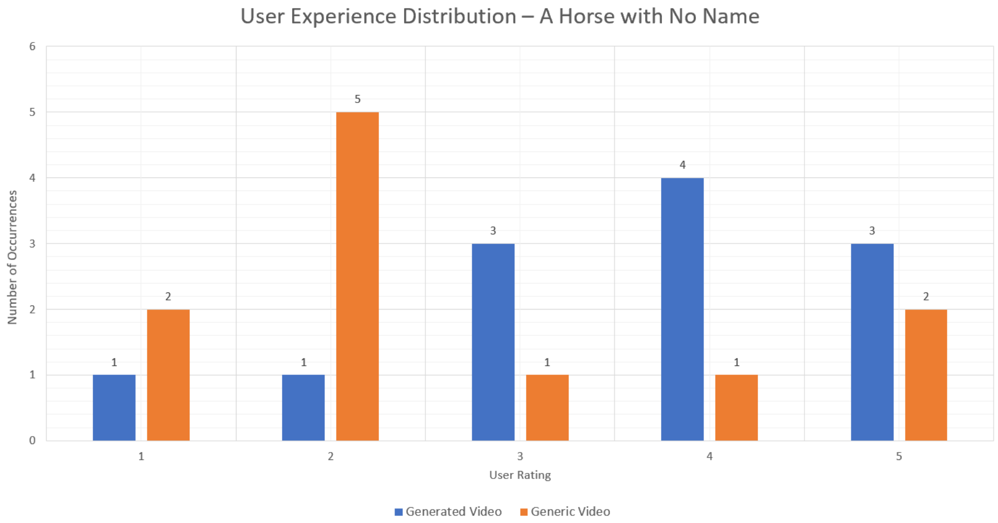


Figure 5.3: *The results of the user experience question referring to the generated video of America's song,* "A Horse With No Name".

Song Name	Video Type	Mean	Median	Variance	p values
A Horse With No Name	Generated	3.5833	4	1.5378	0.061552
	Generic	2.5	2	2.0909	
Roar	Generated	3.25	3.5	2.3864	0.59775
	Generic	3.5833	4	2.2652	
What Makes You Beautiful	Generated	2.9231	3	2.4103	0.23214
	Generic	2.1818	2	1.9636	

Table 5.2: *Table containing the user experience result calculations for each video.*

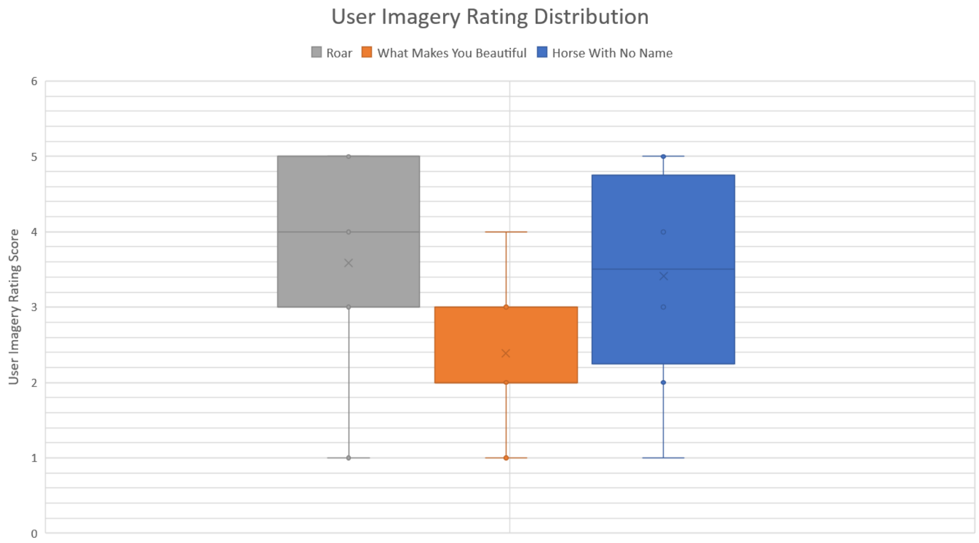


Figure 5.4: Figure representing the distribution of all votes regarding the imagery for each of the generated videos.

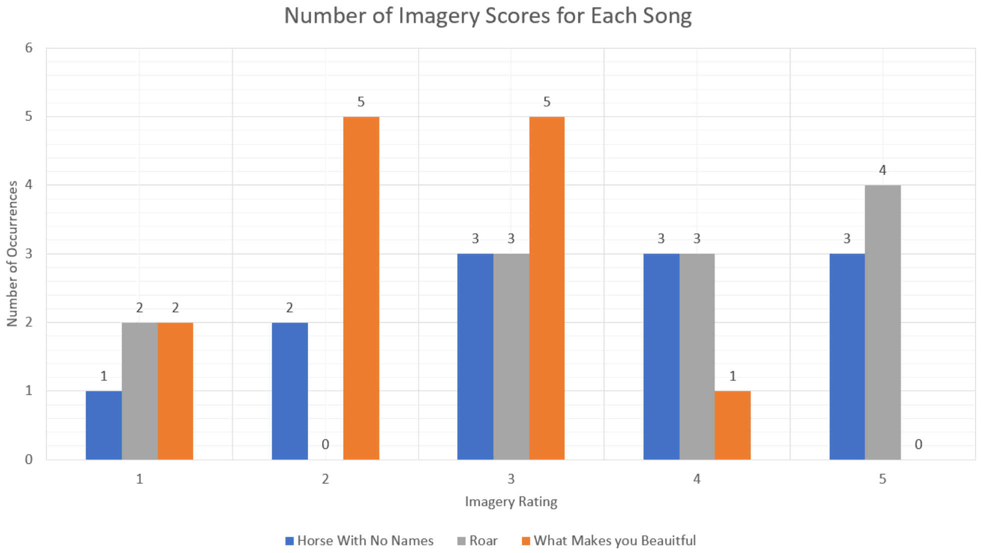


Figure 5.5: Figure representing all the occurrences of all votes regarding the imagery for each of the generated videos, derived from table A.3

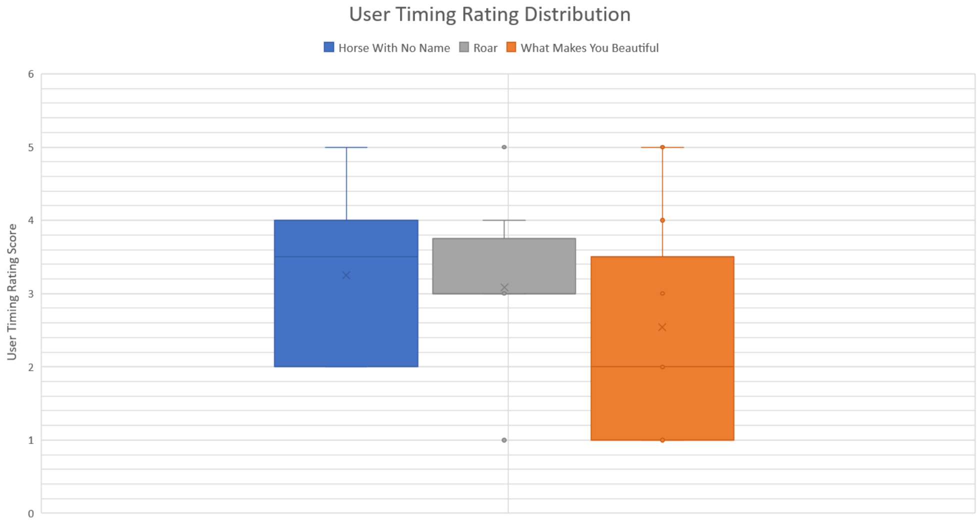


Figure 5.6: Figure representing the distribution of all votes regarding the timing for each of the generated videos.

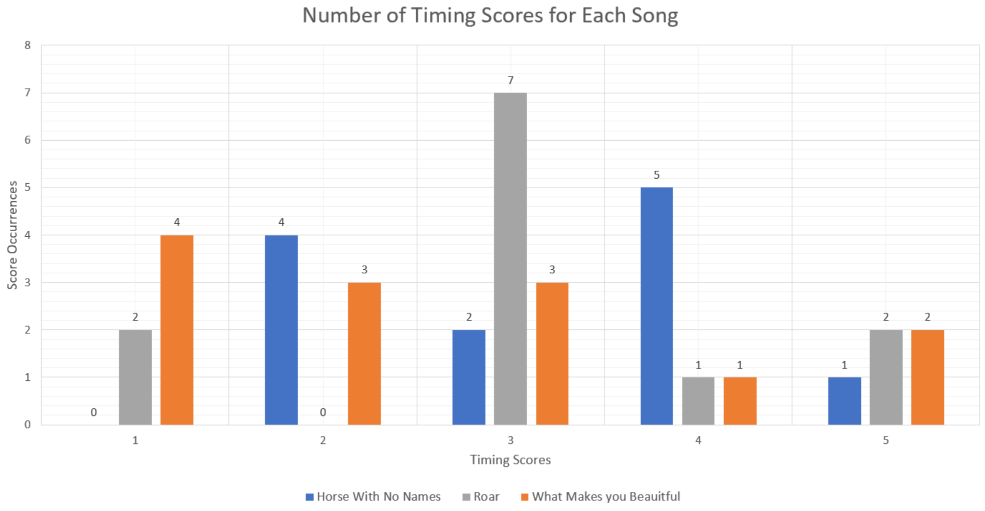


Figure 5.7: Figure representing all the occurrences of all votes regarding the timing for each of the generated videos, derived from table A.2.

6 Conclusion

This project has produced an Automatic Videography system capable of generating music videos with imagery representing the lyrics of a song through various methods, which include Forced Alignment, Lyrical Analysis ,and Caption Extraction.

We have provided an overview of the past technological pressures on the music industry which has shaped and continues to shape the musical landscape. We have also researched the impacts that a system capable of increasing the modality of an audio source has within the field of accessibility studies, and explored the economic effects of the increased prevalence of the user-generated content on streaming platforms. As well as the limitations of state-of-the-art and industry standard speech recognition models when applied to lyrical recognition.

The system was evaluated through exposing surveyed participants to the videos produced by the system to determine if user experience improved, compared to the experience of participants viewing 'still' videos. The results of the survey were insufficient to reject the null hypothesis that user experience would be unaffected through exposure to either type of video. However, the user feed provided allowed for speculation about the reasons within and outside of the system to explain these results. These reasons are worthy of additional research, such as investigating the relationship between user attitudes and their perceptions of system performance on user experience. Users provided insights about the improvements that could be made to the system in future iterations, which would facilitate the increase in prevalence of more accessible content.

A Appendices

Video Type	Song Name	User Ratings				
		1	2	3	4	5
A Horse With No Name	Generated	1	1	3	4	3
	Generic	2	5	1	1	2
Roar	Generated	3	0	3	3	3
	Generic	2	1	1	4	4
What Makes You Beautiful	Generated	3	3	2	2	3
	Generic	5	2	2	1	1

Table A.1: *Table containing all the occurrences of all votes for each video, both generic and generated.*

Video Type	Song Name	Timing Rating				
		1	2	3	4	5
Generated	A Horse With No Name	0	4	2	5	1
Generated	Roar	1	0	7	1	5
Generated	What Makes You Beautiful	4	3	3	1	2

Table A.2: *Table containing all the occurrences of all votes regarding the timing for each of the generated videos.*

Video Type	Song Name	Imagery Rating				
		1	2	3	4	5
Generated	A Horse With No Name	1	2	3	3	3
Generated	Roar	2	0	3	3	4
Generated	What Makes You Beautiful	2	5	5	1	0

Table A.3: *Table containing all the occurrences of all votes regarding the imagery for each of the generated videos.*

6 Bibliography

[1] Academic. Non-lexical vocables in music. https://en-academic.com/dic.nsf/enwiki/7789441, 2022. Accessed: 2022-03-29.

[2] G. Azuaje, K. Liew, E. Epure, S. Yada, S. Wakamiya, and E. Aramaki. Visualyre: Multimodal visualization of lyrics. In *Audio Mostly 2021*, pages 130–134. 2021.

[3] J. Barker, S. Watanabe, E. Vincent, and J. Trmal. The fifth'chime'speech separation and recognition challenge: dataset, task and baselines. *arXiv preprint arXiv:1803.10609*, 2018.

[4] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509: 257–289, 2020.

[5] J. Depoix. youtube-transcript-api. https://pypi.org/project/youtube-transcript-api/, 2021. Accessed: 2022-02-17.

[6] M. Digital, Culture and S. Committee. Economics of music streaming. https://committees.parliament.uk/publications/6739/documents/72525/default/, 2021. Accessed: 2022-02-16.

[7] Django. The web framework for perfectionists with deadlines. https://www.djangoproject.com/, 2005. Accessed: 2022-02-17.

[8] I. Doumanis, D. Economou, G. R. Sim, and S. Porter. The impact of multimodal collaborative virtual environments on learning: A gamified online debate. *Computers & Education*, 130:121–138, 2019.

[9] G. B. Dzhambazov and X. Serra. Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In *Timoney J, Lysaght T, editors. 12th Sound and Music Computing Conference; 2015 jul. 30-ag. 1; Maynooth (Ireland). Maynooth: Music Technology Research Group, Department of Computer Science, Maynooth University; 2015. Oral session 7, Computational musicology and mathematical music theory 1; p. 281-286.* Music Technology Research Group, Department of Computer Science, Maynooth ..., 2015.

[10] S. R. El-Beltagy and A. Rafea. Kp-miner: A keyphrase extraction system for english and arabic documents. *Information systems*, 34(1):132–144, 2009.

[11] J. Frost. Null hypothesis: Definition, rejecting examples. https://statisticsbyjim.com/hypothesis-testing/null-hypothesis/, 2022. Accessed: 2022-03-28.

[12] H. Fujihara and M. Goto. Lyrics-to-audio alignment and its application. In *Dagstuhl Follow-Ups*, volume 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[13] S. Funasawa, H. Ishizaki, K. Hoashi, Y. Takishima, and J. Katto. Automated music slideshow generation using web images based on lyrics. In *ISMIR*, pages 63–68, 2010.

[14] R. Garcia. youtube-dl. https://github.com/ytdl-org/youtube-dl, 2021. Accessed: 2022-02-17.

[15] Genius. Genius. https://genius.com/, 2009. Accessed: 2022-02-17.

[16] Google. Add subtitles and captions. https://support.google.com/youtube/answer/2734796?hl=en-GB#zippy=, 2022. Accessed: 2022-03-26.

[17] Google. Overview of copyright management tools. https://support.google.com/youtube/answer/2807622, 2022. Accessed: 2022-03-03.

[18] G. M. Greco. Accessibility studies: Abuses, misuses and the method of poietic design. In *International Conference on Human-Computer Interaction*, pages 15–27. Springer, 2019.

[19] G. M. Greco. Towards a pedagogy of accessibility: The need for critical learning spaces in media accessibility education and training. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 18, 2019.

[20] G. M. Greco and A. Jankowska. Media accessibility within and beyond audiovisual translation. In *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, pages 57–81. Springer, 2020.

[21] C. Gupta, R. Tong, H. Li, and Y. Wang. Semi-supervised lyrics and solo-singing alignment. In *ISMIR*, pages 600–607, 2018.

[22] C. Gupta, E. Yılmaz, and H. Li. Automatic lyrics alignment and transcription in polyphonic music: Does background music help? In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 496–500. IEEE, 2020.

[23] O. HASLAM. Apple has a team of people transcribing song lyrics for apple music. https://www.imore.com/apple-has-team-people-transcribing-song-lyrics-apple-music, 2019. Accessed: 2022-03-28.

[24] O.-P. Heinisuo. opencv-python. https://pypi.org/project/opencv-python/, 2021. Accessed: 2022-02-17.

[25] J. Huggins. selenium. https://pypi.org/project/selenium/, 2004. Accessed: 2022-02-17.

[26] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[27] M. Lee. pytesseract. https://pypi.org/project/pytesseract/, 2021. Accessed: 2022-02-17.

[28] T. W. Leong and P. C. Wright. Revisiting social practices surrounding music. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 951–960, 2013.

[29] L. A. Liikkanen and A. Salovaara. Music on youtube: User engagement with traditional, user-appropriated and derivative videos. *Computers in Human Behavior*, 50:108–124, 2015.

[30] A. Mesaros and T. Virtanen. Automatic alignment of music audio and lyrics. In *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*, 2008.

[31] A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010:1–11, 2010.

[32] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[33] MIREX. Lyrics alignment results. https://www.music-ir.org/mirex/wiki/2017:Automatic_Lyrics-to-Audio_Alignment_Results, 2017. Accessed: 2022-03-23.

[34] L. of Artificial Intelligence and D. Support. yake. https://github.com/LIAAD/yake, 2021. Accessed: 2022-02-17.

[35] D. of Computer Science. Ethics checklist form. http://www.dcs.gla.ac.uk/~hcp/ethics/projects-form.pdf, 2022. Accessed: 2022-03-28.

[36] O. J. of the European Union. Opinion of the european economic and social committee on 'accessibility as a human right for persons with disabilities'. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52013IE3000&from=GA, 2014. Accessed: 2022-03-19.

[37] W. H. Organization. World report on disability. https://www.who.int/disabilities/world_report/2011/report.pdf, 2011. Accessed: 2022-03-19.

[38] RAILSWARE. Moscow method: How to make the best of prioritization. https://railsware.com/blog/moscow-prioritization/#Why_do_you_need_prioritization, 2022. Accessed: 2022-03-26.

[39] K. Reitz. requests. https://pypi.org/project/requests/, 2022. Accessed: 2022-02-17.

[40] L. Richardson. beautifulsoup4. https://pypi.org/project/beautifulsoup4/, 2021. Accessed: 2022-02-17.

[41] N. F. Ronnie Ghose, Taylor Fox Dahlin. pytube. https://github.com/pytube/pytube, 2022. Accessed: 2022-02-17.

[42] S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20, 2010.

[43] D. Stoller, S. Durand, and S. Ewert. End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–185. IEEE, 2019.

[44] M. M. Tom Lehman, Ilan Zechory. Genius. https://pypi.org/project/lyricsgenius/, 2009. Accessed: 2022-02-17.

[45] P. D. Turney. Learning algorithms for keyphrase extraction. *Information retrieval*, 2(4): 303–336, 2000.

[46] X. Wan and J. Xiao. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, volume 8, pages 855–860, 2008.

[47] Y. Wang, M.-Y. Kan, T. L. Nwe, A. Shenoy, and J. Yin. Lyrically: automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 212–219, 2004.

[48] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. Kea: Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, pages 129–152. IGI global, 2005.

[49] Zulko. moviepy. https://pypi.org/project/moviepy/, 2017. Accessed: 2022-02-17.